

Lecture Notes in Physics

N. D. Hari Dass

Strings to Strings

Yang-Mills Flux Tubes, QCD Strings
and Effective String Theories

 Springer

Lecture Notes in Physics

Founding Editors

Wolf Beiglböck

Jürgen Ehlers

Klaus Hepp

Hans-Arwed Weidenmüller

Volume 1018

Series Editors

Roberta Citro, Salerno, Italy

Peter Hänggi, Augsburg, Germany

Morten Hjorth-Jensen, Oslo, Norway

Maciej Lewenstein, Barcelona, Spain

Luciano Rezzolla, Frankfurt am Main, Germany

Angel Rubio, Hamburg, Germany

Wolfgang Schleich, Ulm, Germany

Stefan Theisen, Potsdam, Germany

James D. Wells, Ann Arbor, MI, USA

Gary P. Zank, Huntsville, AL, USA

The series Lecture Notes in Physics (LNP), founded in 1969, reports new developments in physics research and teaching - quickly and informally, but with a high quality and the explicit aim to summarize and communicate current knowledge in an accessible way. Books published in this series are conceived as bridging material between advanced graduate textbooks and the forefront of research and to serve three purposes:

- to be a compact and modern up-to-date source of reference on a well-defined topic;
- to serve as an accessible introduction to the field to postgraduate students and non-specialist researchers from related areas;
- to be a source of advanced teaching material for specialized seminars, courses and schools.

Both monographs and multi-author volumes will be considered for publication. Edited volumes should however consist of a very limited number of contributions only. Proceedings will not be considered for LNP.

Volumes published in LNP are disseminated both in print and in electronic formats, the electronic archive being available at springerlink.com. The series content is indexed, abstracted and referenced by many abstracting and information services, bibliographic networks, subscription agencies, library networks, and consortia.

Proposals should be sent to a member of the Editorial Board, or directly to the responsible editor at Springer:

Dr Lisa Scalone

lisa.scalone@springernature.com

N. D. Hari Dass

Strings to Strings

Yang-Mills Flux Tubes, QCD Strings
and Effective String Theories

N. D. Hari Dass
Institute of Mathematical Sciences (Retired)
Chennai, India

ISSN 0075-8450 ISSN 1616-6361 (electronic)
Lecture Notes in Physics
ISBN 978-3-031-35357-4 ISBN 978-3-031-35358-1 (eBook)
<https://doi.org/10.1007/978-3-031-35358-1>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland



*This book is dedicated to my mother,
N.S.Venktalakshmi (1926-2021). She would
be enquiring every time I saw her about the
progress of the book, till the last day of her
life.*

Picture credit: Ravi Balasubramanyam

Preface

This book is all about what I would describe as an unprecedented situation in physics where two seemingly orthogonal descriptions of Strong Interactions, namely, String Theory and QCD, eventually reach a fusion in the effective string theory of Yang-Mills flux tubes. Hence, the main title is *Strings to Strings*, and the subtitle is *Yang-Mills Flux Tubes, QCD Strings and Effective String Theories*. In contrast to String Theory, QCD or *Quantum Chromodynamics* is a Relativistic Quantum Field Theory. This is indeed an extraordinary scientific journey. Both these descriptions happened around the same time too, i.e. the early 1970s. It is extraordinary because the idea of hadronic strings was a culmination of a journey whose starting point was the rejection of *Quantum Field Theory* as the correct language for the description of strong interactions, because of the many failures of the Meson Field Theories.

This author was fortunate to have contributed to this journey both through numerical simulations based on Lattice Gauge Theories as well as analytical works in the so-called Effective String Theories done with his collaborators Pushan Majumdar (who passed away at a very early age 2 years ago), Peter Matlock and Yashas Bharadwaj. Significant parts of the important last two chapters are based on these contributions.

The idea of writing this book originated in a suggestion by Prof. B. Ananthanarayan that I write a contribution to the Springer Briefs in Physics. My original aim was to describe the effective string theories of flux tubes as such a Springer Brief. But the book proposal I made turned out to be much more ambitious and Lisa Scalone, Physics Editor at Springer Nature, suggested that I write the book as a Lecture Notes in Physics.

My plan was to explain all the concepts in as much technical detail as possible. I also wanted the book to be a single resource for the wide-ranging ideas and techniques that spanned this journey. This was largely motivated by my own sense of despair while trying to grasp these ideas as a graduate student around the same time that both String Theory and QCD were completed. The major source of that despair was not knowing where to look and how many to look! Therefore, I have taken extraordinary efforts in describing each link in this marvelous chain of developments. This meant going to the original papers and often reworking their presentations to make the contents easily accessible to students. This turned out

to be particularly challenging as even famous original papers had typos as well as errors, not to mention frustratingly different conventions and symbols. The writing has turned out to be deeply personal in the sense that I got to know many leading contributors to this journey personally!

The book is divided into four parts. The first part is non-mathematical but otherwise technically faithful description which traces the historical developments all the way from our conception of Elements to the failures of Mesonic Field Theories in describing nuclear phenomena. The second part, titled *Heisenberg's S-matrix to String Theory* describes, through 14 chapters, the milestones in the development of String Theory. These chapters include comprehensive coverage of non-perturbative relativistic quantum field theories, the Kallen-Lehmann Representation, the Lehmann-Symanzik-Zimmermann formalism in great detail including unitarity and analyticity aspects of it, the Lehmann Ellipses, a long chapter delineating many aspects of Dispersion relations, theory of Complex Angular Momentum and Regge Poles, superconvergence relations and duality, extensive coverage of the famous Veneziano amplitude and its N-point generalizations and a thorough presentation of the operator formalism for Dual Resonance Models. I chose to present these in such detail because of the essential roles they played in the construction of String Theory. This part concludes with the chapter on String theory and its quantization both according to the original work of Goddard, Goldstone, Rebbi and Thorn as well as the one by Arvis of strings with fixed ends. Both these are repeatedly invoked in the crucial chapters of the book.

The third part called *Strings Lost* describes the essentials of QCD, essentials of Lattice Gauge Theory along with a chapter on effective field theories. These form the backbone to the subsequent description of flux tubes both from a theoretical as well as a numerical simulation point of view. The chapter on effective field theories covering effective descriptions of superconductivity as well as of strong interactions is to let the readers appreciate the spirit and technical details of effective string theories later. This chapter also brings out many important connections between superconductivity and strong interaction physics. The chapter on QCD apart from discussing standard details like renormalization group and asymptotic freedom, also has extensive discussion of the static quark-antiquark potential as well as the dual superconductor mechanism for quark confinement.

In the last part called *Strings Regained*, the book has two extensive chapters on the studies of flux tubes in Lattice Gauge Theory and Effective String Theories. In the former, it traces the progress made in the numerical investigations of flux tubes through both the static quark-antiquark potential as well as the profiles of flux tubes. It is these simulations that eventually pointed to the rather precise and remarkable parallel between flux tubes and Bosonic String Theory. In the last chapter, the idea of effective string theories and their application to understand and explain the observed features of flux tubes is explained in great detail for both their static gauge formulation pioneered by Lüscher and Weisz, as well as the conformal gauge formulation of Polchinski and Strominger. The chapter also discusses a powerful Covariant Calculus, developed by the author with Peter Matlock, for

a systematic construction of effective string theories. It also reviews the current status of effective string theories.

I have chosen the layout and sequence of chapters very carefully to blend both pedagogical and research objectives. What is needed in a chapter is thoroughly covered in the previous chapters. I sincerely hope that the readers will benefit from this and also have fun reading though some chapters are necessarily technical.

Even after more than five decades since the inception of QCD, many foundational issues are still not fully understood some of which have persisted from the very early days of nuclear physics. Even the concept of S-matrix which played a major role in the developments described in the book remains a difficult conceptual problem because of quark and gluon confinements. It is hoped that this book provides the inspiration and necessary technical background to researchers wanting to take this exciting journey forward.

The challenges of writing this book have been compounded by the pandemic, as well as my mother's passing away. This has led to frequent revisions of deadlines. I thank Lisa Scalone for putting up with these and for her constant encouragement to go on. Encouragement and support from Murugappan Muthukumar is also greatly appreciated. I also thank Chandra Sekaran Arjunan for prompt help in technical matters. My special thanks are to C. V. Ramamurthy for the drawings, to The American Physical Society, Journal of High Energy Physics, Gunnar Bali, Martin Lüscher and Peter Weisz for permissions to reproduce from earlier publications, and Ravi Balasubramanyam for the use of a picture of my mother.

The pandemic made the paucity of library support much worse, and I am indebted to the Leiden University Library in providing a copy of Kramer's seminal paper on dispersion relations which was very hard to locate (an Italian conference, a paper by a Dutchman and in French!). Paul Pandian, the librarian at IMSc, Chennai, has been of immense help in getting me copies of original papers from a wide variety of sources, often at very short notice. Likewise, Rutvij Bhavsar's help in this regard has also been indispensable. Anil Shaji helped me in getting hold of the paper by Sudarshan and Marshak. Likewise, K. S. Mallesh and Srinivas Prabhu helped in getting hold of John Carson's classic on Electric Circuit Theory. M. Sivakumar gifted me his book on QFT along with many discussions on various aspects of the renormalization group. The help of Kalyan Kumar, Suman and Srinidhi from Tata Institute of Fundamental Research, Hyderabad in setting up a VPN connection to their library has been crucial. Bernard de Wit helped me with the translation of a crucial paper by Ralph Kronig, which was in Dutch. I thank them all wholeheartedly. I am also thankful to Jnanadeva Maharana for an illuminating discussion on the proof of fixed-t dispersion relations. I also acknowledge with thanks useful correspondences with Sumit Das and Bastian Brandt.

I am thankful to V. Ravindran, Director of IMSc, for the hospitality of a month when some important chapters got written. My thanks are also to Justin David of CHEP, Indian Institute of Science, Bangalore, for arranging a visit which also helped in finishing the book.

This book marks a special moment in my long journey in the world of physics. That journey itself owes to many influences; that of my high school teacher H. Anant Rao who made me chose physics, that of many excellent teachers at Delhi University, and, in particular, Prof. A. N. Mitra for my initiation into the world of QFT and Prof. S. N. Biswas into S-matrix theory, my thesis advisor Raymond Sawyer into the world of chiral symmetry and current algebras, Jim Hartle, Bob Sugar and Richard Blankenbecler at UCSB into many aspects of physics and so many friends and collaborators. Highly insightful discussions with Hikaru Kawai, 'tHooft, the late Pierre van Baal, Vikram Vyas and Apoorva Patel have been critical.

The book is a culmination of several years of very difficult technical work. My heartfelt appreciations to my collaborarators Pushan Majumdar, Peter Matlock and Yashas Bharadwaj.

Needless to say, my total preoccupation with the writing of the book has caused immense difficulties to my immediate family and close friends, particularly to my wife Jayanthi and my daughter Shantala. I am indebted to their constant support despite the odds. A constant beacon of support has been my late mother, who, without knowing the physics behind the book, would every day enquire about its progress till her last day. This book is dedicated to her.

Mysore, India
May 2023

N. D. Hari Dass

Contents

Part I The Background

1	Elements to Elementary Particles	3
1.1	Introduction	3
1.2	The Elements	3
1.3	The Atoms	4
1.4	The Elementary Particles	6
1.4.1	The Spin Angular Momentum	7
1.4.2	Statistics and Spin	10
1.5	Quantum Electrodynamics	12
	References	16
2	Radioactivity and Weak Interactions	19
2.1	Radioactivity	19
2.2	Energy Spectrum of β -Electrons	20
2.3	Wolfgang Pauli and the Neutrino	21
2.4	Fermi Theory of Beta Interactions	22
2.4.1	Neutrino Masses	22
2.4.2	Generalizing Fermi Theory	23
2.5	Even More Particles	23
2.6	Intermediate Vector Bosons	24
2.7	The Electroweak Unification	26
	References	27
3	Nuclear Forces, Meson Field Theories and Their Failures	31
3.1	Nuclear Forces: Observational	31
3.2	Yukawa Theory	33
3.3	Meson Field Theories and Their Failures	34
3.4	Experimental Discoveries of the Pions	35
3.5	Meson Theories Post-Pion Discovery	36
	References	38

Part II Heisenberg's S-matrix to String Theory

4	The S-matrix: From Heisenberg Till Now	43
4.1	Introduction	43
4.2	Kramers, Kronig and Analyticity	45
4.3	Connections to Causality	49
4.4	Causality and Analyticity in Non-relativistic Quantum Mechanics	52
4.5	Microcausality	53
	References	56
5	QED: S-Matrix, Causality and Analyticity	59
5.1	Development of QED	59
5.2	Dyson Equivalence Proof	71
5.3	S-Matrix in QED	73
5.4	Causality, Analyticity, and S-Matrix in Non-perturbative RQFT	74
	References	75
6	A Non-perturbative RQFT Primer	77
6.1	QFT-A Particle Perspective	77
6.1.1	Notations and Conventions	78
6.2	Quantum Fields from Particles	78
	References	86
7	The Kallen-Lehmann Representation	87
7.1	Introduction	87
7.2	The Kallen-Lehmann Representation	88
7.2.1	Lessons for Analytic S-Matrix	93
	References	94
8	The Lehmann Symanzik Zimmermann (LSZ) Formalism	95
8.1	The LSZ Reduction Formulae	95
8.1.1	The Retarded Commutator Representation	100
8.1.2	What if No Tricks are Used?	102
8.1.3	Unretarded Commutator Representation	102
8.1.4	Crossing Symmetry	103
	References	105
9	Unitarity and the LSZ Formalism	107
9.1	Introduction	107
9.2	General Considerations	109
9.3	LSZ Formalism and Unitarity	112
9.3.1	Proof of Unitarity in LSZ Formalism	113
	References	116

10	Lehmann Ellipses	117
10.1	Lehmann Ellipses	117
10.1.1	Jost-Lehmann-Dyson Theorem	119
	References	125
11	Dispersion Relations in RQFT	127
11.1	Introduction	127
11.2	Toll's Analysis of the Logical Foundations	128
11.3	Dispersion Relations in QFT: General Considerations	130
11.4	Forward Scattering Dispersion Relations	132
11.4.1	Massless Particle Scattering	132
11.4.2	Massive Particle Scattering: Goldberger Analysis	134
11.4.3	Massive Particle Scattering: Symanzik Analysis	141
11.5	Non-forward Scattering: Salam's Approach	144
11.6	Fixed-t Dispersion Relations: Lehmann and Sommer	151
11.7	Mandelstam Double Spectral Representation	158
	References	160
12	Some Uses and Applications of Analyticity and Dispersion Relations	163
12.1	Low Energy Meson-Nucleon Scattering	163
12.2	Pion Decay	164
12.3	The Froissart, Khuri-Kinoshita Bounds	166
12.3.1	The Pomeranchuk Theorem	167
12.3.2	Pi-Pi Scattering	168
12.3.3	Some Recent Developments	168
12.4	Adler-Weisberger Relations for g_A	169
	References	173
13	In the Land of Complex Angular Momentum	175
13.1	Introduction	175
13.2	Lehmann Ellipses and Partial-wave Analysis	176
13.3	Going Beyond Lehmann Ellipses: Complex Angular Momentum	178
13.3.1	Cases where Angular Momentum can be Complexified	179
13.4	The Sommerfeld-Watson Transform	182
13.5	Regge Poles and Their Properties	184
13.5.1	Bound States and Resonances	185
13.5.2	Regge Asymptotics	186
	References	188

14	Superconvergence Relations, FESR and Duality	189
14.1	Introduction	189
14.2	Superconvergence Relations	189
14.2.1	Igi's Significant Next Step	191
14.3	Finite Energy Sum Rules(FESR)	193
14.3.1	Horn-Schmid Formulation	193
14.3.2	Dolen, Horn, Schmid Elaboration	194
14.3.3	Igi-Matsuda FESR	196
14.4	Alarm Bells Regarding FESR	197
14.4.1	The Mandula-Slansky Work	198
14.4.2	Fujisaki's Work	199
14.5	Concluding Remarks	201
	References	201
15	The Veneziano Formula and the Dual Resonance Model	203
15.1	Introduction	203
15.2	The Veneziano Formula	204
15.2.1	Veneziano's Motivation	204
15.2.2	Important Properties of the Veneziano Formula-I	205
15.2.3	Precise Duality	208
15.2.4	The Issue of Daughters	211
15.3	The Multi-point Function Generalizations	214
15.3.1	The 5-Point Function	214
15.4	The Higher Point Functions	220
15.4.1	The 6-Point Function	220
15.4.2	N-point Function	222
15.4.3	Koba-Nielsen Variables	225
	References	231
16	The Operator Formalism and The Dual Resonance Model	233
16.1	Introduction	233
16.2	Operator Formalism-I	234
16.2.1	Preliminaries	234
16.2.2	Oscillators and N-Point Functions	236
16.2.3	Factorizability and Degeneracies	240
16.3	Operator Formalism-II	242
16.4	Physical States of the Dual Resonance Models	244
16.4.1	Varieties of States of the Dual Model	247
16.4.2	Absence of Ghosts and DDF Construction	248
16.4.3	QED Revisited	249
16.4.4	First Two Excited States	249
16.5	The Shapiro-Virasoro Model	251
	References	253

17	The Birth of String Theory	255
17.1	Introduction	255
17.2	Emergent Strings and Their Actions	256
17.2.1	Nielsen-Susskind Action	256
17.2.2	Nambu-Goto Action	259
17.3	Classical Analysis of the Nambu-Goto Action	260
17.3.1	Light-Cone Parametrization	265
17.3.2	Non-covariant Quantization	268
17.4	Covariant Quantization	269
17.5	The Arvis Quantization	272
17.6	Path Integral Quantizations	275
	References	275

Part III Strings Lost: QCD, The Field Theory of Strong Interactions

18	Effective Field Theories	279
18.1	Introduction	279
18.2	Effective Description of Weak Interactions	280
18.3	Effective Descriptions of Superconductivity	283
18.4	Effective Descriptions of Strong Interactions	288
18.4.1	Group Structure of Chiral Transformations	297
18.4.2	Spontaneous Breaking of Chiral Symmetry	297
18.4.3	Non-linear Realization of Chiral Symmetry	301
18.4.4	Chiral Cancellations	302
18.4.5	Finite Pion Mass	303
18.4.6	Phenomenological Lagrangeans	304
18.4.7	Chiral Perturbation Theory	308
18.4.8	Anomalous Sector	310
	References	311

19	Quantum Chromodynamics (QCD)—A RQFT for Strong Interactions	315
19.1	Introduction	315
19.2	Historical Backgrounds to QCD	316
19.2.1	Strangeness	316
19.2.2	Sakata Model	316
19.2.3	The Eightfold Way	317
19.3	The Quark Models	318
19.3.1	Gell-Mann-Zweig Quark Models	318
19.3.2	The Statistics Difficulties	319

19.3.3	Han-Nambu (HN) Quark Model	319
19.3.4	π^0 -Decay as a Test for Quark Models	320
19.3.5	Colour in Gell-Mann-Zweig Model	321
19.4	Towards Theories of Quark Dynamics	322
19.4.1	Han-Nambu Approach	322
19.4.2	Fritzsch-Gell-Mann Approach	323
19.4.3	Observational Tests for Han-Nambu Model Revisited	325
19.4.4	Deep Inelastic Scattering	325
19.5	Quantum Chromodynamics (QCD)	326
19.5.1	A QED Interlude	326
19.5.2	QCD	327
19.6	Asymptotic Freedom and Perturbation Theory	330
19.6.1	Sliding Scales and Running Couplings	330
19.6.2	The Callan-Symanzik Equation and the Beta Function	330
19.6.3	The Running Couplings	332
19.6.4	The Beta Function	333
19.6.5	RG in QED	335
19.6.6	Improving Perturbation Theory	337
19.6.7	RG in QCD	337
19.6.8	Higher Order Beta Functions	340
19.6.9	More on Asymptotic Freedom	341
19.7	The Static Quark-Antiquark Potential in QCD	343
19.8	Colour Confinement in QCD	344
19.9	QCD and Chiral Symmetry	347
	References	348
20	Essentials of Lattice Gauge Theories (LGT)	351
20.1	Introduction	351
20.2	An Elementary Introduction to Lattice Field Theories	359
20.2.1	The Statistical Continuum Limit	361
20.2.2	Statistical Continuum Limit of the $D = 1$ Example	363
20.3	Gauge Fields on Lattices	365
20.3.1	Abelian Gauge Fields	365
20.4	Non-Abelian Gauge Fields on Lattice	366
20.4.1	Invariants of Non-Abelian Gauge Theories	368
20.4.2	Wegner-Wilson Loops	369
20.4.3	Polyakov Lines (Loops)	370
20.4.4	The Plaquette Action	370
20.4.5	The LGT Path-Integral	371
	References	372

Part IV Strings Regained: From Yang-Mills Flux Tubes to Effective String Theories

21	Lattice Gauge Theory and Yang-Mills Flux Tubes	377
21.1	Introduction	377
21.2	Flux Tube Observables	379
21.2.1	The Static $Q\bar{Q}$ -Potential	379
21.2.2	Confinement Criterion	380
21.2.3	The Polyakov Lines (Loops)	380
21.2.4	The Flux Tube Profile	381
21.3	Creutz's Pioneering Numerical Works	383
21.3.1	Monte Carlo Simulations	384
21.3.2	The Statistical Continuum Limit	388
21.4	Work of Ambjorn, Olesen and Peterson	391
21.5	Flux Profile Studies	392
21.6	Simulations of Lüscher and Weisz	394
21.7	Simulations of Hari Dass and Pushan Majumdar	396
	References	399
22	Flux Tubes and Effective String Theories (EST)	401
22.1	Introduction	401
22.2	Lüscher-Weisz Effective String Theories	403
22.2.1	Leading Order Analysis	404
22.2.2	A Possible Boundary Term	405
22.2.3	Dimension-2 Corrections	406
22.2.4	Open-Closed String Duality	407
22.2.5	Purely Classical Analysis	409
22.3	Polchinski-Strominger(PS) Effective String Theory	411
22.3.1	Leading Order Analysis of PS Effective Actions	412
22.4	PS Effective String Theories for all D	414
22.4.1	Order R^{-3} Corrections to the Spectrum	418
22.4.2	Ground State Momentum Revisited	422
22.5	Covariant Calculus for Effective String Theories	424
22.5.1	Covariant Calculus I: The Nambu-Goto way	425
22.5.2	Covariant Calculus II: The Polyakov Way	427
22.5.3	Weyl Connections and Weyl-Weight Compensators	429
22.6	Gauge Fixing the Covariant Actions	431
22.6.1	The Static Gauge	431
22.6.2	Covariant Calculus I: The Conformal Gauge	431
22.6.3	Covariant Calculus II: The Conformal Gauge	432
22.7	Equivalence of Conformal Gauges	434
22.8	Drummond Actions as Examples	435

22.9	Spectrum of Effective Strings at Even Higher Orders	436
22.9.1	Results by Aharony et al.	436
22.9.2	Alleged Equivalence to Arvis Spectrum To All Orders	438
22.10	Other Important Issues	441
22.10.1	The Excited States	441
22.10.2	AdS-CFT Approaches	443
22.10.3	Thickness of Flux Tubes	444
22.11	Path Integral Quantization of Subcritical Strings	445
22.12	Concluding Remarks	448
	References	449
Index		453

Acronyms

ADM	Arnowitt, Deser, Misner
AdS	Anti-de Sitter
AF	Asymptotic Freedom
BCS	Bardeen, Cooper, Schrieffer
BCH	Baker, Campbell, Hausdorff
BJ	Bjorken-Drell
BR	Barton
BST	Bosonic String Theory
CO	Carbon Monoxide
CFT	Conformal Field Theory
CVC	Conserved Vector Current
DDF	Di Vecchia, Del Giudice, Fubini
DRM	Dual Resonance Models
EFT	Effective Field Theory
EOM	Equation Of Motion
EST	Effective String Theories
ETCR	Equal Time Commutation Relation
FESR	Finite Energy Sum Rules
GGRT	Goddard, Goldstone, Rebbi, Thorne
GL	Ginzburg-Landau
GLL	Ginzburg-Landau-London
GR	General Relativity
HM	Hari Dass Matlock
HN	Han Nambu
JOSA	Journal Optical Society of America
KN	Koba Nielsen
LGT	Lattice Gauge Theory
LLE	Large Lehmann Ellipse
LSZ	Lehmann Symanzik Zimmerman
LW	Lüscher-Weisz
MS	Minimal Subtraction
NG	Nambu-Goto
NRQM	Non-relativistic Quantum Mechanics
OPE	Operator Product Expansion

PCAC	Partially Conserved Axial Current
PDG	Particle Data Group
PL	Polyakov-Liouville
PS	Polchinski-Strominger
QCD	Quantum Chromo-Dynamics
QED	Quantum Electro-Dynamics
QFT	Quantum Field Theory
QM	Quantum Mechanics
$Q\bar{Q}$	Quark-Antiquark
RG	Renormalization Group
RQFT	Relativistic Quantum Field Theory
SLE	Small Lehmann Ellipse
SV	Shapiro-Virasoro
UHP	Upper Half Plane

Part I

The Background



1.1 Introduction

It is not too often (in fact the author can not think of even once) in the history of science that two *radically different* theories were put forward to explain the same phenomena, with each one running into conceptual and technical difficulties of a very serious kind, with each ultimately triumphing over those difficulties, and miraculously, for each to be discovered in the other! The phenomenon we are referring to is that of *Strong Interactions*, more popularly known as *Nuclear Forces*, and the radically different theories were (i) *Theory of hadronic strings* and (ii) *Quantum Chromodynamics*. The latter is also an example of what are called *Relativistic Quantum Field Theories*. In this book, it is my goal to explain this rather remarkable occurrence in the history of modern physics as thoroughly as I can. It is understandably very technical, involving conceptual and technical aspects of both String Theory as well as Quantum Field Theory. However, in the first three chapters, I explain the genesis of all the relevant ideas, from the intuitive concepts of elements, in a non-technical manner that should be understandable to a very wide audience.

1.2 The Elements

Man's curiosity about the surrounding world appears to have prodded him (her) to conjecture a hypothetical, simpler world from which the actual world could be worked out as a consequence. A most creative and powerful manifestation of this process is the idea of *Elements*! One can think of the elements as the basic building blocks out of which all *others* are built. The concept of nature, and of elements, that had the strongest influence on modern science was that of the Greeks. Empedocles, as early as 450 BC, had listed earth, air, fire and water as the elements. All these early

ideas of elements led to little progress till seventeenth century AD. They lacked any precise criterion for what could be considered an element, or otherwise. A conceptual breakthrough was made by Robert Boyle (1627–1691), who sharpened the notion of elements to be substances that could not be “broken down” to simpler substances. By this criterion, none of the substances considered as elements earlier could really be accepted as one. For example, water is decomposable into oxygen and hydrogen (first demonstrated by Nicholson and Carlisle in 1800 using a chemical battery). But there is no known way of further “breaking down” oxygen and hydrogen. Care should be taken to understand that the breaking down we are talking about is *chemical*, or *electrochemical*. Otherwise, one can bombard the nuclei of oxygen till they break down into protons and neutrons!

This is a good point to pause to point out that scientific method acquires its power through the clarity of the conjectures made. The earlier characterizations of elements were rather vague as criteria, while Robert Boyle’s refinements gave rise to immediate and much progress, as we shall see shortly. Much before the dissociation of water in 1800, Lavoisier by 1789 had successfully isolated several elements. By 1808, Humphrey Davy made impressive additions by isolating the elements potassium, sodium, barium and strontium. Soon after, the number of elements started proliferating.

1.3 The Atoms

It is the atomic hypothesis, that all substances are made up of *atoms* that made the idea of elements a powerful driving force of modern physics. The idea of atoms is again very old, almost as old as that of elements itself. Democritus in the fourth century BC and Leucippus in the fifth century BC are the most famous from Greece. Kanada from India, of the fourth to sixth century BC, is credited with having proposed an atomic hypothesis. The crux of these very early atomic “theories” was that all substances are made of very small entities called atoms which were said to be indestructible. But again, as in the case of elements, these ideas were not concrete enough to have had any lasting impact. Just as Robert Boyle’s refinements were what was required to make real progress, it was John Dalton (1766–1864)’s rather precisely formulated laws that made the atomic hypothesis amenable to further analysis and progress.

Let us list the essence of Dalton’s laws: **D1:** All elements are made up of extremely small particles called *atoms*. **D2:** Atoms of a given element are *identical* in size, mass and other properties; atoms of different elements *differ* in their size, mass, and other properties. **D3:** Atoms can not be subdivided, created or destroyed. **D4:** Atoms of different elements combine in simple *whole number ratios* to form *chemical compounds*.

Dalton formulated these laws around 1803. The precise wordings of D2 and D4 will be found to be crucial later on. D4 is also known as Dalton’s law of multiple proportions. It is the most powerful argument in support of atoms. In an earlier version of this Dalton had taken the ratio to be 1, i.e. he had assumed that one atom of an element only combines with one atom of any other element it combines with. It is also

called the *law of simple proportions*. A most powerful and remarkable application of D4, demonstrated by Dalton himself, was in experimentally determining the so-called *Relative Atomic Masses*. In simple language, these are nothing but the ratios of the various atomic masses to the mass of a hydrogen atom (almost).

To see how it is done, let us consider the case of x gms of carbon burning with y gms of oxygen to produce $z = x + y$ gms of CO . These are of course not exact but hold to great accuracy. x gms of carbon will have $\frac{x}{m(\text{C})}$ atoms of carbon (here $m(\text{C})$ is the mass of a carbon atom in gms). Under Dalton's law of simple proportions (applicable in this case), these will combine with an *equal* number of oxygen atoms with total mass $y = \frac{x}{m(\text{C})} \cdot m(\text{O})$. Since x, y and z are all directly *measurable*, the ratio $\frac{m(\text{C})}{m(\text{O})}$ is directly measurable. Depending on the conventions used for relative masses, $rm(\text{A}) = 12 \cdot \frac{m(\text{A})}{m(\text{C})}$ if defined relative to Carbon-12, or, $rm(\text{A}) = 16 \cdot \frac{m(\text{A})}{m(\text{O})}$, if defined relative to Oxygen-16. From a physicist's elementary particle physics perspective, it is most natural to define relative atomic masses w.r.t hydrogen, as the nucleus of a hydrogen atom, proton, is an elementary particle. But from a chemists point of view, it's a lot more useful to define relative to an element that easily combines with many other elements.

Dalton had prepared an elaborate table of relative masses. While many of them were very close to integers with a high accuracy, there were glaring exceptions too. Chlorine had a value of 35.5, neon had 20.2, etc. These numbers have not changed substantially from those days. It is important to note that there is no conflict between these non-integer values and Dalton's laws, in particular D2. They just imply that the masses of chlorine atoms and neon atoms, belonging to different elements, are just different, and more specifically are in the ratio of 35.5:20.2.

At this point comes another atomic theory, put forward by William Prout (1785–1850) in 1815. According to his hypothesis, *Atomic weights (masses) of elements are whole number multiples of the atomic weight of hydrogen*. In other words, relative atomic weights must all be integers. The aforementioned data of chlorine and neon contradict this. So, does it mean that Prout hypothesis was clearly wrong and to be discarded? The ultimate answer to this question is another great triumph of the scientific method which had far-reaching consequences for our understanding of the structure of matter.

Before that, some important developments need to be mentioned. In 1869 Mendeleev came up with his famous periodic table. In that, the elements were arranged according to their atomic weights. Maxwell formulated his famous laws unifying electricity and magnetism in 1861–62. This was a case par excellence of the so-called *Field Theories*. One may wonder if it has any place in this discussion of elementary particles; it indeed has a very eminent place! With the discovery of the *dual nature* that eventually led to a complete elucidation of quantum phenomena, it led to the concept of the *photon* as a particle. On top of that, the fully quantum description of the Maxwell theory led to the concept of *Relativistic Quantum Field Theories* without which the present-day understanding of elementary particles would have been inconceivable. Rutherford, Geiger and Marsden carried out the now famous gold-foil experiments around 1908–1911 [1, 2], suggesting the idea of very compact atomic nuclei where the net positive charges of the atom are concentrated. An impor-

tant new concept, the atomic number, was introduced by Moseley around 1913 [3], based on his extensive studies of X-rays from various elements, and identified it with nuclear charge. This resulted in the extremely useful classification of nuclei in terms of A the atomic weight (neglecting the small masses of electrons), and Z the atomic number.

1.4 The Elementary Particles

Electrons (cathode rays) were discovered by Julius Plücker in 1859, but it was only the systematic studies by J. J. Thomson which established them as particles with the well-defined charge-to-mass ratio in 1897. Protons had been discovered even earlier in 1886 as *Canal Rays* by Eugen Goldstein, and put on a firmer foundation by Wilhelm Wien in 1898. Year 1913 also saw the birth of the Bohr Model of the atom, a singularly eventful development in the history of atoms. This was extended substantially by Sommerfeld. The Bohr–Sommerfeld model of the atom had many successes, but equally many failures.

Both the electrons and protons had been discovered using the humble *Crookes Tube*. Both these fundamental discoveries made use of an elementary application of Maxwell’s electromagnetism to the motion of charges under simultaneous influence of electric and magnetic fields to determine the charge-to-mass ratio $\frac{q}{m}$. Now comes Francis Aston whose genius at instrumentation leads to a major breakthrough. While there are ways to determine atomic numbers through scattering of Rutherford type, or through spectroscopic means, direct determination of atomic mass was something else. Aston improved upon J. J. Thomson’s apparatus to determine this directly, again by exploiting motion under simultaneous electric and magnetic fields. The basic idea was that particles with different $\frac{q}{m}$ would traverse different trajectories. (For a very account of Aston’s works and their tremendous impact on science see [4]).

In 1909, Aston [4] obtained the first glimpses that neon atoms came in two varieties! One type had a mass of 20, while the other had 22. Their natural abundances were such that the *average* atomic mass of neon was 20.2. Both the atomic masses were indeed integers. This means Prout was actually right and Dalton wrong. This points out the pitfalls of relying on the scientific method too unquestioningly. Aston confirmed what Soddy had observed earlier in the context of radioactivity that elements with same chemical properties could have different physical properties, something Dalton’s laws could not have allowed. These were called *Isotopes*.

Soddy had conjectured the element Lead to have two isotopes, one occurring as the end-product of radioactive decay and the other occurring naturally. The confirmation of this, by Horovitz [5], took painstaking and life-threatening work much like the pioneering works of Marie and Pierre Curie. Horovitz, despite her brilliance, could not get a job in academia due to the prevailing male dominance and died tragically in a Nazi concentration camp.

But the real implications of Aston’s discoveries went far beyond Dalton versus Prout. Since the isotopes 20 and 22 are chemically identical, they must have the same atomic number, and hence the same number of protons. Since their masses are again

integers, there must be other particles in the nucleus with masses nearly same as the protons. Yet they have to be electrically *neutral*. Rutherford anticipated these as early as 1920 in one of his Bakerian lectures [6], naming them *neutrons*. They were experimentally discovered only in 1932 by Chadwick [7].

With this, we have arrived at the elementary particles constituting atoms including their nuclei. The impact and importance of neutrons for any understanding of nuclear physics can not be emphasized sufficiently. They have revolutionized our understanding of astrophysics too, i.e. *Neutron Stars*.

By 1925, Bohr, Heisenberg and Schrödinger had fully developed Quantum Mechanics. Though Albert Einstein had given the world his *Special Theory of Relativity* in 1905 itself, its deep implications for the world of elementary particles had to await the *Dirac Equation* in 1928 [8]. It is worth mentioning at this point that Aston's mass spectrograph has been made so precise that the tiny deviations in the atomic masses from integer values can be understood in terms of Einstein's celebrated $E = mc^2$! The Dirac equation naturally accounted for spin and predicted the existence of *anti-particles*.

1.4.1 The Spin Angular Momentum

The concept of spin had already played a major role in both atomic and nuclear physics even prior to the Dirac equation. Let us briefly narrate the exciting story of spin. In a by now-celebrated experiment, Gerlach and Stern [9] first provided inklings about the idea of spin, even though at the time of the experiment its connections to spin were hardly understood. In this experiment, a beam of silver atoms was passed through an inhomogeneous magnetic field. Classically, one would have expected a force on it depending on the orientation of its magnetic moment making the atoms follow a continuous family of trajectories. Instead, it was observed that the atoms followed one of a well-separated discrete set of trajectories.

More than the Stern–Gerlach experiment, it was the Zeeman Effect, particularly the Anomalous Zeeman Effect that brought the electron spin out of its closet. In 1896, Zeeman [10] observed that when atoms were placed in a magnetic field, the spectral lines split into three. This is called the normal Zeeman effect and a classical explanation could be readily provided on the basis of Lorentz's "theory of electrons" of 1892. Two years after Zeeman's pioneering work, Preston [11] observed a much more complex pattern of spectral lines, now called the *Anomalous Zeeman Effect*. Despite the best efforts neither could one find an explanation in terms of the classical electron theory nor from the Bohr–Sommerfeld model of the quantum atom. Nevertheless, many regularities in these patterns had been found by Preston himself along with Lande, Paschen, etc.

It was a brilliant suggestion of Pauli [12] that paved the way for a resolution of this challenge, along with truly revolutionary implications for all of physics till today. One should read Pauli's Nobel lecture [13] for a proper understanding of what led Pauli to this remarkably bold and innovative idea. At that time, Pauli was also preoccupied with the observed lengths of periods 2, 8, 18, 32... in the chemical

elements, for which the Bohr–Sommerfeld theory had no explanation. In fact, the Bohr–Sommerfeld theory had no explanation for why all the electrons in the atom did not just occupy the ground state, and why the closed-shell picture of Bohr and Sommerfeld was even necessary. Even if one accepted the closed-shell idea, the periodicities would be 1, 4, 9, 16... off by a factor of 2 from the observed patterns.

In 1925, Heisenberg with his *Matrix Mechanics*, and Schrödinger with his *Wave Mechanics* succeeded in creating the quantum theory, which could be successfully applied to the atom. Nevertheless, the above-mentioned difficulties could not be tackled by the new quantum mechanics. It was Pauli's sheer genius that found a common resolution to this vexing problem along with the other vexing problem of the anomalous Zeeman Effect.

The first breakthrough Pauli achieved was in noting that an additional *two-valued* quantum number, over and above the quantum numbers n, l, m_l of the Bohr–Sommerfeld atom could explain the pattern of spectral lines of the Anomalous Zeeman effect. Pauli described this as “A new quantum property, a particular two-valuedness which can not be described from a classical point of view”. The factor of 2 could also account for the observed periodicities of chemical elements if the concept of closed shells was extended to include the new quantum numbers too. Still, the question as to why all electrons did not simply occupy the ground state remained. It turns out Lande [14] too had discovered the additional quantum numbers which he called the *inner quantum number* already in 1923, and Stoner [15] had shown how to get the observed periodicities of 2, 8, 18,... already in 1924. In fact, in his Nobel lecture, Pauli explicitly refers to the impact of Stoner's ideas on his thinking. He also devotes considerable attention to the details of Stoner's work in his famous 1925 paper and how his work leads to a generalization of Stoner's work.

Pauli's second breakthrough was even more spectacular. Announced in the same year 1925, this was the famous *Exclusion Principle* [16]. To quote Pauli verbatim, “There can not be two or more equivalent electrons in an atom, for which in strong fields the values of all four quantum numbers coincide. If an electron exists in an atom for which all of these numbers have definite values (in external fields), then this state is ‘occupied’”. Pauli's quote is in German and what is given here is the author's best compromise between a translation by Kaplan [17], and another by Duck and Sudarshan [18]. Both these translations suffer from inadequacies.

With the Exclusion Principle, the way the electrons occupy various energy levels is elegantly resolved, and the outstanding problem of why all electrons are not found in the ground state simply disappears. Though discovered in the context of atomic physics, this principle has played a central and powerful role in every aspect of modern physics that includes nuclear physics, condensed matter physics, astrophysics (white dwarfs and neutron stars, for example), elementary particle physics at the highest realized energies and even in String Theory. A much more powerful statement of the exclusion principle shall emerge after we have discussed Spin.

Around the same time as Pauli's paper, in January 1925, when Ralph Kronig heard of the double-valued new quantum numbers of Pauli, he proceeded to construct a model whereby the electrons are “spinning” around their “axis” [19]. The primary motivation for Kronig was two-fold; to think of the new quantum numbers as being

associated with angular momentum of a new kind (of a value of $\frac{\hbar}{2}$), and to ascribe to the electrons a new magnetic moment to quantitatively account for the splittings observed in anomalous Zeeman effect. While the double-valuedness of the new quantum numbers can accommodate the additional levels observed, they can not determine the various splittings. Kronig found that ascribing a new magnetic moment of 1 Bohr magneton ($g_S = 2$, in modern parlance), could explain the anomalous Zeeman splitting.

A classically rotating charge distribution only develops a magnetic moment of $\frac{1}{2}$ Bohr magneton ($g_L = 1$, also true quantum-mechanically). This is already a pointer to something rather unusual with the spinning electron idea. But Kronig ran into a difficulty with the so-called *fine structure*. In 1887, Michelson and Morley (yes, the same duo whose null result played an important role in the context of special relativity theory) had shown that what had been earlier thought of as a single spectral line of hydrogen was actually split into many, with splittings about ten thousandth of the normal splittings. This was possible only because of the incredible increase in the resolving powers. Kronig sought to explain this accurately observed fine structure in alkali atoms (doublet D-line) on the basis of a coupling between the magnetic moment due to the orbital motion of the electrons with his proposed magnetic moment due to spin. This interaction came to be known later as the *spin-orbit coupling*. He obtained results twice as large as what was observed. He could not explain this discrepancy.

Pauli was highly dismissive of Kronig's ideas. His specific objections were (i) an electron with so much angular momentum has to rotate with surface velocities far exceeding the velocity of light and (ii) failure to account for fine structure. Kronig abandoned his proposal, now seen in retrospect to be correct [19]. His velocity of light objections appears to be a crude estimate not based on any special relativistic calculations. Any calculation consistent with special relativity can never lead to such conclusions. In fact, immediately afterwards Eddington pointed out that special relativity does not lead to any such restrictions. The fine structure criticism was more subtle, but here too Pauli had been too quick in his criticism, as we shall see shortly.

Around autumn 1925, Goudsmit and Uhlenbeck [20,21] independently came up with essentially the same ideas as those of Kronig, i.e. $g_S = 2$ for spin angular momentum $\frac{\hbar}{2}$ to explain the anomalous Zeeman splittings, but with the same factor two discrepancy in accounting for fine structure. In February 1926, 23-year-old Llewellyn Thomas pointed out that the spin-orbit interaction requires a special relativistic correction. Rather remarkably, the effect of this correction is to effectively halve the spin-orbit interaction! This came to be known subsequently as *Thomas Precession* [22]. It is necessary to understand relativistic precession of every kind, including gravitational! With this correction, all objections to both Kronig, and, Goudsmit–Uhlenbeck spin proposals disappeared. Even Pauli came around to accepting the spin idea. What is most extraordinary in this context is that Pauli was considered among the leading authorities on special relativity, and yet, he failed to see this important consequence. In the light of all this, the credit for electron-spin should go equally to Kronig, Goudsmit and Uhlenbeck.

In 1924, Pauli wrote another foundational paper [23]. It was about how the magnetic moments of *Nuclei*, if they existed, could affect the atomic levels. This is now

called the *Hyperfine Structure*. The effect is $\frac{m_e}{m_p}$ or about thousandth of the fine structure splittings. It is extraordinarily small, and Michelson observed them for the first time in thallium and mercury, again making use of his remarkable interferometer. With this work, Pauli brought nuclear physics and atomic physics together. In 1930, Fermi [24] gave a detailed theoretical treatment of the hyperfine interaction. The atomic hyperfine splittings became a reliable tool for determining nuclear spins (for a detailed introduction see Bethe and Bacher [25]).

1.4.2 Statistics and Spin

With the development of Quantum Theory, the Pauli Exclusion Principle got reformulated as a statement of the so-called *Statistics* of quantum particles. For a system of many identical particles, the statistics was *Bose–Einstein* if the wavefunction was *symmetric* under the exchange of two particles or *Fermi–Dirac* if it was *antisymmetric* under an exchange. The former was discovered in 1924 by Bose [26], for light quanta (photons), and the latter, independently by Fermi and Dirac [27] for electrons. Particles obeying Bose–Einstein statistics came to be called *Bosons*, and those obeying Fermi–Dirac statistics as *Fermions*. Fermions obey Pauli-Exclusion Principle while Bosons do not. The idea of statistics can immediately be applied to systems of many identical particles. While any number of Bosons still behaves as a Boson, *Even* number of Fermions behave as Bosons, and *odd* number of Fermions behave as Fermions. At this point, it is worth recalling that electrons which are Fermions have spin $\frac{1}{2}$ (in units of $\hbar$). In 1931, Raman and Bhagavantam determined photon spin to be 1 [28].

The Spin-Statistics Connection: The above-mentioned connection between spin and statistics, i.e. half-integral spin and fermions, and integral spins and bosons hold for a system with an arbitrary number of particles. The key to this observation is the fact that quantum orbital angular momentum is always *integral*.¹ For a system of N -bosons, the total angular momentum is still integral, and the statistics Bose–Einstein. This is again true for an even number of fermions. For an odd number of fermions, the total angular momentum is half-integral and the statistics is Fermi–Dirac. This connection between spin and statistics, though so simple to state, is really very hard to prove. See Duck and Sudarshan for an exhaustive account.

Both statistics and spin are directly observable (See Bethe and Bacher [25] for a good discussion). One direct way is to look at the rotational bands of diatomic molecules. The intensities of the spectral lines give information about the nuclear spins too. Another accurate determination of both nuclear spin and nuclear magnetic moments is through the hyperfine splittings.

¹ The quantum theory of angular momentum was worked out in 1925 itself by Born, Heisenberg, and Jordan [29] as part of the mathematical formalism of quantum mechanics. Though at that time spin angular momentum was not known, they still concluded that angular momentum can take both integral and half-integral values. This was because they had not insisted on the single-valuedness of wavefunctions.

Proton and Neutron Spin determinations: The hyperfine splitting in hydrogen is too small to have been determined spectroscopically. The 21-cm spectral line, so central to radio astronomy, was observed only in 1951. Dennison [30], had used the low-temperature behaviour of hydrogen molecules to deduce that it is as if there were two separate gases with weight factors in the ratio 1:3. From this, he deduced that the spin of the proton is 1. The two species are called Para and Ortho hydrogen.

Soon after the discovery of neutrons by Chadwick in 1932, the statistics of deuterons was determined to be Bose. If deuterons were to be treated as being constituted by one proton and one neutron, as surmised by Rutherford as early as 1920 [6], it would immediately follow that neutrons too must obey Fermi–Dirac statistics, with half-integral spin. The spin of deuterons had been determined to be 1. This still allowed, as per the quantum theory of angular momentum, values $\frac{3}{2}$, $\frac{1}{2}$ for the neutron spin. Schwinger and Teller [31] had shown that neutron scattering on ortho versus para hydrogen would give useful information on *nuclear forces*. Later, Schwinger showed that the same scattering data could also determine the spin of the neutron. According to him, for neutrons around 100 K, the ratio of ortho to para scattering will be much larger than 1 if neutron spin was $\frac{1}{2}$, while the ratio will be close to unity if the neutron spin was $\frac{3}{2}$ [32]. The experiments of Otto Stern and collaborators [33], as also that of Dunning and collaborators [34] had clearly shown the ratio to be much larger than one. Thus, Schwinger concluded that evidence strongly favoured neutron spin to be $\frac{1}{2}$.

Models of Nuclei before the discovery of neutrons: In the above, we used the nuclear model wherein nuclei are made up of protons and neutrons. This is what Rutherford had envisaged as early as 1920, but it was only a conjecture. Only in 1932 did Heisenberg [35] and Iwanenko [36] put forward this model after extensive experimental investigations in nuclear physics along with powerful theoretical arguments.

It is worthwhile to take a look at the most favoured picture of the nucleus before this. Before the discovery of the neutrons, the general thinking was to model the nucleus with the then-known elementary particles, i.e. protons and electrons. This was a conservative but well motivated attempt. Protons being nearly 2000 times more massive than electrons, this picture required as many protons in the nucleus as the atomic mass (weight) A . Then the number of electrons required to reduce the nuclear charge would be $A - Z$. The total number of Fermions in the nucleus would be $2A - Z$.

From the extensive experimental studies, the following important correlations had been found: (i) nuclei with even atomic weight had integral spins, while those with odd atomic weights had half-integral spins. This was called the **Odd-Even effect** (see [25]); (ii) independently, and consistent with this, it had also been observed that even A nuclei were Bosons, while odd A nuclei were Fermions. Both these correlations were independent of the atomic charge Z . These findings immediately contradicted the proton-electron model of the nucleus. For example, Rasetti [37] had determined that both deuteron and N^{14} were bosons while the proton-electron model would have predicted them to be Fermions because of their odd atomic numbers of 1 and 7, respectively. In general, it would be Z that would determine statistics according to

this model. With the discovery of neutrons which had the same mass as protons to high accuracy, the atomic mass A could now be interpreted as the number of protons plus the number of neutrons. This picture could explain the observed correlations in a simple manner.

Curiously, Kronig, the first discoverer of electron spin, had, in a paper in 1926 [38], had given an argument against electron spin and the then proposed magnetic moment equivalent to $g_S = 2$, which could alternately have been used against the electron-proton model of the nucleus. He had argued that, with an odd number of electrons in nuclei, one would end up with hyperfine splittings (as per Pauli's 1924 ideas) some thousand times larger than the observed values. He wanted this to be an argument against $g_S = 2$ for electrons. On the other hand, the same argument would work against electrons in nuclei, if their spin and magnetic moment had been correct!

1.5 Quantum Electrodynamics

Bohr's atomic theory was born in 1913, and it only dealt with circular orbits. Little later, in 1916, Arnold Sommerfeld showed, in the first of a series of three papers, how to include elliptical orbits. This required the introduction of three quantum numbers n, l, m_l in place of a single quantum number of the Bohr theory. But the energy levels of the Sommerfeld theory were exactly the same as those of the Bohr theory, but with different degeneracies (levels with the same energy). The degeneracies could not account for the observed periodicities 2, 8, 18, ... of chemistry. However, there was only a factor of 2 amiss. The observed fine structure, discussed earlier, could not also be explained. In the second of the papers, Sommerfeld gave a remarkable, fully *relativistic* treatment according to which the energy levels were dependent on two quantum numbers n, j . This remarkable formula could explain a large amount of the fine structure. In fact, it was in this paper that Sommerfeld coined the phrase *fine structure constant* for $\frac{e^2}{\hbar c}$ with the approximate numerical value of $\frac{1}{137}$.

Sommerfeld's analysis was based on the now defunct concept of orbits. For Heisenberg, what paved the way for Quantum Mechanics was the renunciation of the orbits. Schrödinger's wave mechanics also did away with them. The next crucial development was Dirac's relativistic electron theory of 1928. The relativistic wave equation for electrons proposed by Dirac, the famous Dirac equation [8], incorporated electron spin in an elegant and natural manner. It predicted the spin to be $\frac{\hbar}{2}$ and also $g_S = 2$, vindicating the pioneering ideas of Kronig, Goudsmit and Uhlenbeck. In the same paper, Dirac applied his equation to determine the spectrum of the hydrogen atom. He stopped short of completing the analysis. That was done by Gordon [39] (23 Feb 1928; Dirac's paper was submitted on 2 Jan 1928), and independently, by Darwin (6 March 1928) [40]. The paper in *Zeitschrift f. Physik* had a one-line abstract "Die Theorie ergibt genau de Sommerfeld's Feinstrukturformel" translating to "Dirac theory gives exactly the same formula as Sommerfeld's". This was, and still is, rather remarkable! Biedenharn [41] has analysed this *Sommerfeld Puzzle* in considerable detail and has ascribed the perfect agreement to the same underlying symmetries of the two different situations. It is worth recalling that Schrödinger had

first analysed the hydrogen atom from a relativistic theory without spin (the Klein–Gordon equation). On finding that it gave a formula for fine structure that disagreed with experiments, he did not publish it, and instead went on to his non-relativistic wave mechanics. Biedenharn analyses this case also.

The other, totally unexpected, consequence of the Dirac theory was the prediction of *Antiparticles*! This arose out of what appeared to be an embarrassment for the Dirac theory, namely, the existence of *negative energy* solutions. Dirac reinterpreted this by requiring all negative energy states to be completely filled in the ground state of the quantum vacuum. However, a photon, for example, could be absorbed by a negative energy electron in the *Dirac Sea* and be lifted into an electron of positive energy. That would necessarily leave a *hole* in the Dirac Sea that could be characterized as a particle with positive energy and a positive charge. This was the *anti-electron*. In keeping with the scientific conservatism of the times, Dirac in 1930 tried identifying this with the proton [42], one of the already known particles with positive charge and spin $\frac{1}{2}$. Apart from the large difference in the masses of protons and electrons, the measured value of the magnetic moment of protons was far from the Dirac value. Dirac thought that future developments would resolve these difficulties. But Hermann Weyl, also around the same time, purely on the basis of symmetry arguments, had shown mathematically that the antielectron mass must be the same as the electron mass [43]. A simple but persuasive physics argument against the proton interpretation had been put forward by Oppenheimer (also in 1930) [44]. He argued that if the anti-electron were indeed a proton, hydrogen atom would be unstable against decay into photons, with an extremely small lifetime of about 10^{-10} s! The consensus on the anti-electron as a new particle rather distinct from the proton was fast converging. On 2 August 1932, C.D. Anderson discovered the new particle, which was named *Positron* [45]. The radical new feature of the world of elementary particles was that each particle comes with its own antiparticle. The antiparticle of a photon was itself.

Almost 16 years later, in 1949, Richard Feynman gave an entirely new and novel interpretation of positrons that removed the artificial asymmetry in the Dirac sea picture. Based on the *Green's Function* approach to differential equations, Feynman sought to remove the difficulties with negative energy states by requiring the Green's function for the Dirac equation that would propagate only the positive energy states *forward* in time. Then, there is no more freedom left with regard to propagation *backwards* in time. In non-relativistic theory, or in the relativistic Maxwell theory of electromagnetism, one fixes Green's function for past to vanish, on grounds of causality. So, in the Feynman theory, negative energy solutions only propagate backwards in time, and, as Feynman showed, are indistinguishable from positive energy, positive charge particles propagating forward in time, i.e. Positrons [46]. This work of Feynman does not require his famous *Path Integral Formulation* [47]. The Italian physicist Ettore Majorana had, as early as 1937, voiced his dissatisfaction with the inherent asymmetry in the hole-theoretic formulation of Dirac. In his paper [48], he had come up with a symmetric formulation without any negative energy states. Around the same time, he had also advocated the idea that nuclei consist of neutrons and protons [49], though Iwanenko and Heisenberg had reached that conclusion

already 5 years earlier as already mentioned. In 1933, Majorana also proposed many important ideas regarding the nuclear forces [50].

In non-relativistic quantum theory of both Heisenberg as well as Schrödinger, there is a puzzle with regard to radiating atoms. Even the excited states of the atom being *stationary*, they should not decay at all. The resolution of this lies in the recognition that the electromagnetic field should be quantized too and that the stationary states of the quantal atom are not stationary when both the atom and the field are treated together.

The beginning of Quantum Electrodynamics was in Dirac's 1927 paper, almost immediately after the advent of Quantum Mechanics, but almost a year before the Dirac equation. In this paper, he laid out the basic techniques for a quantum treatment of the electromagnetic field. The main idea behind this paper is the observation that the electromagnetic field is mathematically equivalent to an (infinite) collection of Harmonic Oscillators. The quantization of both the Maxwell field and the Dirac field in interaction turned out to be extremely challenging both technically and conceptually. Some of the daunting technical difficulties were the recurring infinities in most calculations, manifest relativistic covariance, gauge invariance, etc. What started with Dirac in 1927 culminated in the creation of QED by 1948–49 by Tomonaga, Schwinger, Feynman and Dyson. In the intervening period, there were many important works by Dirac, Fermi, Fock, Podolsky, Heisenberg, Weisskopf, Jordan, Stueckelberg, Pais, Sakata, Pauli, Kramers, Heitler and many more (see the extensive works of Darrigol). It is beyond the scope of this narrative to do even partial justice to these pioneers.

What broke the ice was a remarkable meeting of minds in Shelter Island during 1–4 June 1947. The objective of the meeting was to take stock of what could be best described as an impasse. What eventually turned the tide were two remarkable, and crucial, experimental observations, both carried out just prior to the Shelter Island meeting (though both were submitted for publication after the meeting). In order to appreciate the extreme importance of these experiments, let us recall the theoretical status at that point.

Both in the Bohr–Sommerfeld theory as well as in the quantum theories of Heisenberg and Schrödinger of the hydrogen-like spectra, there were a large number of degeneracies. For example, for $n = 2$, the four levels of $2s$, $2p$ all had the same energy. Including spin, as discussed above, the degree of degeneracy of the $n = 2$ level was 8. As per the Dirac theory, these eight levels split into two groups of four degenerate levels each; four belonging to $2p_{3/2}$ and the other four made up of two belonging to $2s_{1/2}$ and two to $2p_{1/2}$, which were still degenerate. The other quantity of interest is the magnetic moment of the electron, which was $g_S = 2$ as per Dirac theory.

The two experiments that occupied the attention at Shelter Islands were (i) that by Lamb and Retherford that had shown that the $2s_{1/2}$ and $2p_{1/2}$ levels were not degenerate but split by a very tiny amount of about 1057 MHz. This came to be known as the *Lamb Shift* and (ii) a precise measurement of the magnetic moment of the electron by Kusch and Foley [51] which revealed that g_S deviated from the Dirac value of 2 by about one part in a thousand. Both indicated a breakdown of the Dirac theory.

Hans Bethe reasoned that the Lamb Shift could, in part, be due to the shift in the electron mass due to its interaction with the quantized electromagnetic field. He had followed Kramer's idea of *mass renormalization* to give a non-relativistic treatment whose numerical value agreed rather well with data. The basic idea of mass renormalization in this context was that the mass of the electron gets shifted due to its interaction with the quantized radiation field, and this happens both to free and bound electrons. In the case of the free electron, the only observable is the sum total of the mass of the electron in the absence of such an interaction (called bare mass), and the said shift. But the main difficulty turned out to be that this calculated shift for both free and bound electrons turned out to be infinite. Bethe used the difference between the shift for the bound case and the free case, to get a finite value.

Going from Bethe's preliminary calculation to a fully relativistic calculation, free of the meaningless infinities in all observable quantities like mass, charge, magnetic moments, energy levels of atoms and scattering amplitudes turned out to be immensely hard. This was the achievement of Tomonaga, Schwinger and Feynman. Apart from manifest relativistic covariance, they showed that the problems of infinities occur for both the charge and mass of the electrons. Once these are fixed through the so-called renormalization, all other observables are free of divergences. Theories satisfying this are called *renormalisable*. The modern view of field theories and renormalization has been dramatically influenced by the ideas of Kenneth Wilson (Schwinger and Feynman had also similar views). As per this, all field theories are to be viewed as *effective theories*, with a specified *cutoff*. The pros and cons of this "pragmatic approach" to physics will not be elaborated further. Dyson demonstrated the equivalence of the three apparently different approaches. He also showed that the renormalization procedure worked to all orders in perturbation theory. Another important contribution of Dyson was the construction of a S-matrix as envisaged by Heisenberg on the basis of the theories of Feynman, Schwinger and Tomonaga, albeit in a perturbative sense.

Triumph came when Schwinger calculated the correction to g_S to be $\frac{\alpha}{2\pi}$, in excellent agreement with the measured values of Foley and Kusch [51]. The Lamb Shift had been calculated in the "old fashioned" hole theory by Kroll and Lamb, by French and Weisskopf and by Fukuda, Miyamoto and Tomonaga. The theoretical value came to be 1057.70 MHz as against the experimental value of 1057.77 MHz. Feynman and Schwinger were the first to calculate Lamb Shift correctly on the basis of the relativistic quantum electrodynamics, now known as QED. In the new language of relativistic quantum field theory, the Lamb Shift calculation had to deal with an effect called *vacuum polarization*. Though its actual contribution to the Lamb Shift is rather small compared to the contribution from the Kramers–Bethe self-energy, conceptually, it was a paradigm shift. The physical basis of this effect is that the electromagnetic field does not just represent photons, but superposition of photon states and electron-positron pair states.

Apart from the amazing agreements with experiments, QED heralded the new age of the *Relativistic Quantum Field Theories* and they laid down the conceptual framework as well as calculational techniques. The reader is encouraged to refer to S. S. Schweber's *QED And The Men Who Made it* [52] for a gripping account of the development of QED.

References

1. H. Geiger, E. Marsden, Proc. R. Soc. Lond. A **82**, 495 (1908)
2. E. Rutherford, Lond. Edinb. Dublin Philos. Mag. Ser. 6 **21**(125), 669 (1911)
3. H.G.J. Moseley, Phil. Mag. 6th Ser. **26**, 1024 (1913)
4. G. Squires, Francis Aston and the mass spectrograph. J. Chem. Soc. Dalton Trans. 3893 (1998)
5. M. Rayner-Canham, G. Rayner-Canham, Stefanie Horovitz: a crucial role in the discovery of isotopes, in *Women in their Element*. World Scientific Publishers
6. E. Rutherford, Proc. R. Soc. A **97**(686), 374 (1920)
7. J. Chadwick, Proc. R. Soc. A **136**, 692 (1932)
8. P.A.M. Dirac, Proc. R. Soc. Lond. A **117**(778), 610 (1928)
9. W. Gerlach, O. Stern, Zeit. f. Phys. **9**, 349 (1922)
10. P. Zeeman, Proc. R. Acad. Amst. **5**(181), 242 (1896)
11. T. Preston, *The Scientific Transactions of the Royal Dublin Society*, 2nd Series 6 (1898), p. 385
12. W. Pauli, Z. f. Phys. **31**, 373 (1925)
13. W. Pauli, *Nobel Lectures, Physics* (Elsevier, Amsterdam, 1964), pp. 1942–1962
14. A. Lande, Z. f. Phys. **16**, 391 (1923)
15. E.C. Stoner, Phil. Mag. **48**, 719 (1924)
16. W. Pauli, Z. f. Phys. **31**, 765 (1925)
17. I.G. Kaplan, *Pauli Exclusion Principle and its Theoretical Foundation*. [arXiv:1902.00499](https://arxiv.org/abs/1902.00499) [quant-ph]
18. I. Duck, E.C.G. Sudarshan, *Pauli and the Spin-Statistics Theorem* (World Scientific Publishers, 1997)
19. E.D. Commins, Ann. Rev. Nucl. Part. Sci. **62**, 133–157 (2012)
20. S.A. Goudsmit, G.E. Uhlenbeck, Naturwissenschaften **13**, 953 (1925)
21. S.A. Goudsmit, G.E. Uhlenbeck, Nature **117**, 264 (1926)
22. L.H. Thomas, Nature **117**, 514 (1926)
23. W. Pauli, Naturewissenschaften **12**, 741 (1924)
24. E. Fermi, Z. f. Phys. **60**, 320 (1930)
25. H.A. Bethe, R.F. Bacher, Rev. Mod. Phys. **8**, 82 (1936)
26. S.N. Bose, Z. f. Phys. **26**, 178 (1928)
27. E. Fermi, Rendiconti Lincei **3**, 145 (1926); P.A.M. Dirac, Proc. R. Soc. A **112**, 661 (1926)
28. C.V. Raman, S. Bhagavantam, Indian J. Phys. **6**, 353 (1931)
29. M. Born, W. Heisenberg, P. Jordan, Z. f. Phys. **35**, 557 (1925)
30. D. Dennison, Proc. R. Soc. A **115**, 483 (1927)
31. J. Schwinger, E. Teller, Phys. Rev. **52**, 286 (1937)
32. J. Schwinger, Phys. Rev. **52**, 1250 (1957)
33. J. Halpern, I. Easterman, O.C. Simpson, O. Stern, Phys. Rev. **52**, 142 (1937)
34. J.R. Dunning, J.M. Manley, H.J. Hoge, F.G. Brickwedde, Phys. Rev. **52**, 1076 (1937)
35. W. Heisenberg, Z. f. Phys. **77**, 1 (1932)
36. D. Iwanenko, Nature **129**, 798 (1932)
37. F. Rasetti, Z. f. Phys. **61**, 598 (1930)
38. R. Kronig, Nature **117**, 550 (1926)
39. W. Gordon, Z. f. Phys. **48**, 11–14 (1928)
40. C.G. Darwin, Proc. R. Soc. Lond. A **118**, 654 (1928)

41. L.C. Biedenharn, Found. Phys. **13**(1), 13–34 (1983)
42. P.A.M. Dirac, Proc. R. Soc. A **126**, 801 (1930)
43. H. Weyl, *Gruppentheorie Und Quantummechanik*, 2nd edn (1930)
44. R. Oppenheimer, Phys. Rev. **35**, 562–563 (1930)
45. C.D. Anderson, Phys. Rev. **43**(6), 491–494 (1933)
46. R.P. Feynman, Phys. Rev. **76**, 749 (1949)
47. L.M. Brown (ed.), *Feynman's Thesis - A New Approach to Quantum Theory* (World Scientific Publishers, 2005)
48. E. Majorana, Nuovo Cimento **5**, 171 (1937)
49. E. Majorana, Nuovo Cimento **14**, 171 (1937)
50. E. Majorana, Kerntheorie. Z. f. Phy. **82**, 137 (1933)
51. P. Kusch, H.M. Foley, Phys. Rev. **73**, 412 (1948)
52. S.S. Schweber, *QED and the Men Who Made it* (Princeton University Press, 1994)

2.1 Radioactivity

Henri Becquerel discovered *radioactivity* in 1896 [1]. He showed that uranium emits “rays” that can affect photographic plates. Shortly before, Wilhelm Röntgen had discovered X-rays. Becquerel’s discovery was immediately followed by Marie Curie’s work showing similar “radiation” was also emitted by thorium [2, 3]. She pioneered the use of measuring the electrical conductivity of air exposed to these rays. She also coined the word *Radioactivity*. The name is confusing as there is no connection of any kind with radio waves. *Radius* is the latin name for “rays”. Marie and Pierre Curie also discovered new elements *Polonium* [4] and *Radium*. Marie Curie intuitively felt that radioactivity was a phenomenon at the “atomic level”, though there was hardly anything at that time to support such a view.

In 1899, Rutherford performed a crucial experiment that showed that radioactivity was of two fundamentally different types, called α , β -rays by him [5]. A year later, in 1900, Paul Villard discovered a third type in his studies on Radium [6]. Later, Rutherford called it γ -rays. The α -radioactivity is a quantum mechanical effect called *tunnelling*, while γ -radioactivity is the result of electromagnetic transitions between excited nuclei, much like the situation in atomic transitions.

In 1900, Becquerel measured the $\frac{q}{m}$ ratio for the β -rays and established they were indeed electrons [7]. This was a very important step in the understanding of the radioactivity phenomena. Next year, Rutherford and Soddy [8] experimentally showed that there was *transmutation of elements* during radioactive decays, except the γ -decays. In 1911, Geiger, Marsden and Rutherford carried out their famous gold-foil scattering experiments [9, 10]. These experiments showed that the positive charge of the atom was concentrated in a very small region, called the nucleus from then on. As mentioned earlier, already in his Bakerian lectures of 1920 [11], Rutherford had

anticipated the Neutrons and had envisaged the nucleus to be composed of protons and neutrons, as confirmed by the nuclear models of Heisenberg [12] and Iwanenko [13] in 1932, also the year of the discovery of neutrons by Chadwick [14].

2.2 Energy Spectrum of β -Electrons

Now, we come to an extremely important topic in the development of the theory of radioactivity; this is the nature of the energy spectrum, or the energy distribution of the electrons emitted in β -decay. This may appear to be going into a lot of details, but as we shall see later, a proper understanding of this played a crucial role. The complexity of the issue can be judged by the fact that it took more than 20 years for the best experimentalists and theoreticians to come to a consensus on this.

Becquerel [7] in 1900 and Kaufmann [15] in 1902 had claimed that β -electrons are emitted with a distribution of velocities. Later on, Becquerel went along with the view that beta-electrons are mono-energetic. In their careful study of the problem during 1908–10, Hahn and Meitner [16] had concluded that, at source, the beta-electrons were monoenergetic but, on detection, there was broadening due to other sources. Meitner, even as late as 1922 maintained this stand [17, 18]. She gave various justifications citing the monoenergetic nature of gamma rays, and that a continuous spectrum was unlikely in quantum theory. Quite clearly, line spectrum had been observed in conjunction with beta-decay. Marie Curie's assistant Jan Kazimierz Danyasz had himself observed several lines as early as 1913 [19]. Rutherford had made the interesting point that photographic methods tended to exaggerate lines against diffuse backgrounds. Chadwick and Ellis [20] had argued in favour of the continuous spectrum. In 1925, Ellis and Wooster countered many of Meitner's conclusions. They had pioneered a new technique of total absorption towards this end. They delivered the definitive word on this complex issue in 1927, and the verdict was the continuous spectra of beta-rays [21, 22].

What about the observed line spectra then? Were they wrong? They were eventually understood as due to the phenomena of *internal conversion*. In this, a potential gamma ray from a decaying nucleus is absorbed by a K-shell electron leading to an emitted electron with definite energy. The hole in the K-shell is subsequently filled up by a nearby electron. Ironically, the experimental evidence for internal conversion was provided by Ellis and Wooster themselves [23, 24]. Swiles [25] was among the first to attempt a theory of the internal conversion process. This is the same Bertha Jeffreys, who along with her husband produced that monumental work on Mathematical Physics. Casimir and Hulme [26, 27] improved Bertha Swirles analysis. Finally, Mott and Taylor gave a correct theory of internal conversion [28]. A fascinating account of this important phase is covered in [29].

2.3 Wolfgang Pauli and the Neutrino

A continuous spectrum posed grave challenges. If the nucleus before, the nucleus after and the electron were the only parties to the decay process, only a line spectrum would be possible. This was also Meitners's reasoning. This reasoning is of course based on the cherished principle of energy conservation. Therefore, one possible explanation would be that energy is not conserved exactly in each beta-decay, but still leaving room for a conservation on average. Even Niels Bohr did not consider this too far-fetched. But Ellis and Wooster rejected this possibility in their 1925 paper. Another possibility was that a neutral particle was emitted along with the electron. That would clearly be compatible with a continuous spectrum for the electron alone. This possibility, which eventually turned out to be the correct one, was not seriously pursued till much later by Pauli in 1930. With the known particles of that time, a likely candidate for such a neutral particle would be a photon (gamma ray). While that could certainly explain the continuous spectrum of the beta-electrons, it would run into other difficulties, as we shall see shortly.

If the only parties to the beta-decay were the initial and final nuclei, and the emitted electron, two difficulties, though related, arise. In beta-decay, the atomic weight A of the nuclei remains the same, though the atomic number Z increases by one unit. From our earlier discussions, this would imply that the statistics of the final nucleus must be the same as that of the initial nucleus; on the other hand, electrons being half-integer spin particles, the spin of the final nucleus would be integral if the initial nucleus has half-integral spin, and the spin of the final nucleus would be half-integral if the spin of the original nucleus were integral. Either way, the statistics of the final nucleus would be different from the statistics of the initial nucleus, and one will be led into a contradiction. Even though the scenario of a gamma ray emitted additionally would solve the continuous spectrum problem, it would still suffer from this statistics and spin problem.

The breakthrough came in late 1930 from Pauli's suggestion of a *neutrino*, which he had originally called a neutron. He made this proposal which radically transformed all of physics, astrophysics and cosmology, in the form of a letter [30] addressed to "Dear Radioactive Ladies and Gentlemen" on 4 December 1930, with the objective of "saving the exchange theorem" (meaning the spin-statistics connection). Quoting Pauli's letter "...could exist in nuclei electrically neutral particles, that I wish to call neutrons, which have spin $1/2$ and obey exclusion principle, and which differ from the light quanta in that they do not travel with velocity of light". What Pauli really meant was that in beta decay, the conjectured particles are emitted along with the electrons, thus solving in one go both the continuous spectrum problem, as well as the spin-statistics problems. Pauli justified this unusual "letter" format as according to him the idea was "too speculative for publication".

2.4 Fermi Theory of Beta Interactions

The idea did not gain much currency till Enrico Fermi made use of it in his ground-breaking work on the theory of beta-decay in 1934 [31]. In that, he renamed Pauli's conjectured particles to *neutrino* ("a little neutron" in Italian). In Fermi theory, the nucleus undergoing beta-decay makes a transition to another nucleus by emitting the "electron-neutrino field" in close analogy with electromagnetism where a charge particle makes transitions from one state to another by emitting the electromagnetic field. The exchanged field in electromagnetism is a Bose field, and Fermi's "electron-neutrino field" is also a Bose field, even though the electron and neutrino fields by themselves are Fermi fields. Inspired by the *current-current* interaction in electromagnetism, Fermi wrote down a Hamiltonian for beta-decay that was a product of the four fermion fields of proton, neutron, electron and neutrino. He wrote this as a contact interaction to reflect the expectation that the range of this interaction is very short. The particular form of the Hamiltonian Fermi wrote down was what one would get for an electrodynamics-like vector current interaction in the limit of non-relativistic protons and neutrons; Fermi brings out the vector nature of the current in the form of the "bilinear" made out of the neutrino and electron fields.

The field-theoretic Hamiltonian could automatically handle radioactive decays with positron emission, electron capture, etc., without introducing additional parameters. With this rather simple and elegant formulation, Fermi could calculate the lifetimes of beta-decay processes, as well as the continuous spectrum of the beta-rays. He introduced what came to be known later as the "Fermi coupling constant" $G_F \simeq \frac{10^{-5}}{m_p^2}$, which he determined after comparing his predictions with data.

2.4.1 Neutrino Masses

Fermi devoted an important part of the paper to the neutrino mass question. As we have already noted, Pauli in his famous letter had explicitly thought of massive neutrinos. Fermi kept the mass of the neutrino as an open question to be decided empirically. He pointed out that the end-point of the beta-electron spectrum is particularly sensitive to the mass of the neutrino. Decades later, it is this technique applied to the tritium beta decay that is one of the most promising laboratory determinations of (electron) neutrino mass. Fermi estimated the mass of the neutrino to be a very small fraction of the electron mass. Later on, it is this tiny mass that prompted many, like Landau [32], to think of massless neutrinos. Pontecorvo was among the first to emphasize that there were no compelling theoretical reasons for the masslessness of neutrinos [33]. It is only with the advent of the V-A theory for weak interactions as proposed by Sudarshan and Marshak [34], and, Feynman and Gell-Mann [35], that a massless neutrino was preferred on theoretical grounds. The experimental discovery of neutrinos by Reines and Cowan [36] was the crowning glory of this chapter in science.

2.4.2 Generalizing Fermi Theory

It was soon pointed out by many that the Fermi Hamiltonian was not the most general. On the experimental side, the Fermi Hamiltonian only accounted for transitions where the nuclear spins did not change. An important improvement came from the works of Gamow and Teller [37], which could accommodate transitions where nuclear spins changed too. But that could only come if, in addition to Fermi's vector current theory, axial currents also contributed. Subsequently, the most general Hamiltonian involving scalar, pseudoscalar, vector, axial-vector and tensor exchanges were considered. This marked the most confusing phase in the development of weak interactions. A revolutionary breakthrough took place with the suggestion of Lee and Yang [38] that there was no evidence for assuming parity conservation in beta-decay, and the subsequent brilliant experimental demonstration by Wu [39] that in fact parity is maximally violated in beta-decay. At the same time, Garwin, Lederman and Weinrich [40] had shown that parity was violated in muon-decay. The same year, the relevant currents were proposed to be actually V-A, as mentioned in the last para.

2.5 Even More Particles

For the major aspects of our book, weak interactions do not play any significant role. We shall be content with merely mentioning the high points in that direction. This will neither be chronological nor in logical sequence. In terms of particles, *muons*, similar to electrons in many respects except for their mass which is nearly 200 times the electron mass, were discovered by Anderson and Neddermeyer [41]. This was really a bolt from the blue as it was just around the time (1935) that Yukawa had proposed the Pi-mesons as the particles responsible for *Nuclear Forces* [42] (lot more on this, shortly). Initially, the Muons were confused for the pions. In intriguing experiments, Bose and Choudhury [43] observed particles with masses clustering around both currently accepted muon and pion masses. In 1940, Williams and Roberts [44] experimentally demonstrated the decay of muons with a high energy electron as among the decay products. They, influenced by Yukawa, had conjectured that muons decay into electrons and a neutrino. In his original work, Yukawa proposed that his mesons were not only carriers of the strong nuclear force but also of beta-radioactivity. For the latter, he had argued pions would decay into an electron and a neutrino. So, Williams and Roberts had (erroneously) argued that this decay was another evidence for the observed mesotrons to be the Yukawa particles. Next year (1941), Rasetti [45] measured the lifetime of muons for the first time. This was considerably improved by the careful expt. of Conversi, Piccioni and Pancini [46]. It took another 7 years before Steinberger [47] in 1948 showed (experimentally) that the electrons in muon decay were not monoenergetic, which meant that the final state had at least three particles. This leads to the recognition that muons were also Fermions. In the meantime, Yukawa's mesons, called Pi-mesons were discovered by Lattes, Occhialini and Powell [48] (charged pions), and by Bjorklund and collaborators (neutral pions) [49]. It is an irony that Yukawa's bold proposal would be upturned

by the discovery of the Muons the very next year. The Yukawa conjecture would have yielded different couplings to nuclear beta decay and muon decay. Much later, the charged pions were seen to decay predominantly into muons and neutrinos. This explains why the identification of the mesotron (subsequently muons) with Yukawa's mesons (pions) by Williams and Roberts was incorrect. In any case, the Yukawa idea would have ascribed to the weak currents a scalar-pseudoscalar nature and not the vector-axial-vector current (V-A) [34,35].

The Fermi theory of beta-interactions, with various refinements that followed, could account for all these mesonic decays very well. The coupling in all of them was more or less still G_F . Thus, the weak interactions too displayed *Universality* in much the same way Gravitation did. For example, the same current-current interaction with one current made up by muons and neutrinos, and the other by electrons and neutrinos, could explain the muon decay. At this point came another deep revelation about neutrinos: that the (anti)neutrino in neutron beta-decay is a different type of (anti)neutrino that accompanies the decay of negatively charged pion into a muon. The former got to be named electron-neutrino ν_e , while the latter was called muon-neutrino ν_μ . This was anticipated by Schwinger [50] on purely theoretical grounds (of conserved lepton number). Experimentally, Lokanathan and Steinberger [51] had shown that the decay of a muon into an electron and photon was highly forbidden. This was also analysed by Feinberg [52]. The two neutrino hypothesis got its brilliant confirmation by an experiment in 1962 at the Brookhaven Lab by Danby and collaborators [53]. This was the first experiment to be performed with reactor neutrinos and was first proposed by Pontecorvo [33]. The neutrinos from π^+ decay were directed against a target and it was found that predominantly only μ^+ were observed. If there was only one kind of neutrino, one should have seen both electrons and muons equally often. Finally, Perl et al. [54] discovered another particle τ , with a mass nearly twice that of a proton. Yet, it was closer to electrons and muons than protons or neutrons in that it participated only in weak and electromagnetic interactions. It too came with its own neutrino, now called tau-neutrino ν_τ . Again, its decays are mediated by weak interactions. Now the nomenclature has changed w.r.t to electrons, muons, tau and their associated neutrinos. They are called *leptons*.

2.6 Intermediate Vector Bosons

Before Yukawa's seminal paper in 1935 [42] about massive mesons as carriers of the forces between nucleons (the collective name for protons and neutrons), the only known case of force carriers were the photons mediating electromagnetic interactions, and photons were massless, reflecting the *infinite* range of electromagnetic forces. Yukawa's was the first instance of a carrier, i.e. mesons, which were *massive*, reflecting this time the finite range of nuclear forces as established by a number of studies in nuclear physics. The complementarity between the range of a force and the mass of its carrier follows from quantum mechanics. Soon after the Fermi theory of beta-interactions (1934), and Yukawa's meson theory (1935), Oskar Klein carried Yukawa's idea to build a theory of beta-interactions. He argued, in this great

pioneering work, that the carriers should be massive vector bosons [55]. The vector nature was to be in accord with Fermi's Hamiltonian. That aspect would survive even after later developments identified the currents to be V-A. Klein even incorporated Heisenberg's concept of *Isotopic Spin* according to which protons and neutrons were to be put on par, as the two components of an $\text{Isospin-}\frac{1}{2}$. In fact, he went far beyond that; he built a field theory of *Gauged Isospin Rotations*. Almost 16 years later, Yang and Mills [56] rediscovered such field theories, and are now known as Yang–Mills fields. The charged part of Klein fields was massive, while the neutral field associated with photons was massless. Klein had to treat both electrons and neutrinos as nearly massless. Klein also did not quantize the vector boson fields.

The idea of massive vector bosons as mediating weak interactions was revived by Schwinger [50]. He too introduced only two charged vector bosons in addition to the photon. The need for such massive vector bosons was also hinted at by Sudarshan and Marshak [34], as well as by Feynman and Gell-Mann [35], by 1957 too. By that time, Glashow had already realized that quantization of massive vector bosons is problematic, and in particular, may lack renormalizability. In 1969, Reiff and Veltman [57] explicitly demonstrated the non-renormalizability of massive Yang–Mills theories. The way the mass problem was overcome is a truly splendid part of the history of modern physics. The key to this was the idea of *spontaneous symmetry breaking*. Heisenberg was the first to investigate this issue in connection with his works on Ferromagnetism [58], barely a few years after the founding of Quantum Mechanics in 1925. In the context of *Superconductivity*, this concept was brought to prominence by Anderson and Nambu [59]. Jeffrey Goldstone provided a general field theoretic understanding [60]. The punch-line was that whenever continuous symmetries (like rotation) are spontaneously broken in the sense that even though the dynamical laws are governed by such symmetries, the ground states are not, there will be associated massless particles. These are called Nambu–Goldstone bosons, and examples are phonons, magnons, rotons, etc., Heisenberg had conjectured that Yukawa's pions are the (nearly) massless particles arising out of spontaneous breaking of isotopic spin symmetry. Nambu also identified the pions as Nambu–Goldstone bosons arising out of the spontaneous breaking of chiral symmetry. The latter is the correct interpretation as it would make the pions pseudoscalar, while Heisenberg's interpretation would only yield scalar pions.

Returning to the mass problem with the intermediate vector bosons of weak interactions, it was found by Englert and Brout [61], Higgs [62], Guralnik, Hagen and Kibble [63] (and perhaps by more) that upon introducing massless gauge fields into the system with spontaneously broken global symmetries (more technically described as “gauging” the relevant global symmetries), the Nambu–Goldstone bosons disappeared. But they did so at the expense of making the gauge vector bosons *massive* (the total number of degrees of freedom can not change.). This came to be known as the *Higgs Effect*. Both Anderson and Nambu (in that order) had already realized that the *Meissner Effect*, whereby magnetic fields are expelled from superconductors, is indeed a manifestation of the Higgs Effect. Some prefer to call it Anderson–Higgs Effect, but as mentioned, many contributed independently.

2.7 The Electroweak Unification

Now we are ready to put various pieces of the Jig-Saw puzzle together, or rather, Salam and Ward, Glashow and Weinberg were the ones ready to do so. The vector bosons of weak-interaction theory were to be like the Yang–Mills–Klein gauge fields, but with their masses generated by the Anderson–Nambu–Higgs–Brout–Englert–Guralnik–Hagen–Kibble mechanism. Bludman [64] was the first to propose an electroweak unification based on SU(2) only. For somewhat technical reasons, Glashow had argued for an additional *neutral vector boson*, also getting its mass through the Higgs effect. The papers of Glashow [65], Salam and Ward [66], appeared in 1959, while the electroweak theory of Leptons was given by Weinberg [67] in 1967. In that work, Weinberg constructed a theory of the vector bosons and the scalar field now called the Higgs Field interacting with left-handed electrons and electron-neutrinos. The coupling to muons and muon-neutrinos was to be just a replication of the coupling to electrons and electron-neutrinos. The coupling to the strongly interacting protons and neutrons was much more subtle. At that time, the symmetry group of hadrons was still thought to be some broken version of Gell-Mann’s SU(3), and it was not obvious how to include that. By that time the Gell-Mann–Zweig *Quark Model*, already introduced in 1964 [68,69] (more on it later, as it led to *Quantum Chromodynamics* (QCD), the currently accepted model of strong interactions [70]), was beginning to be taken seriously and there too there were three quarks, u, d, s to reckon with. A fourth quark had been conjectured by Bjorken and Glashow on grounds of lepton-hadron symmetry as early as 1964 [71] when Gell-Mann and Zweig had proposed their quark model, but there wasn’t any evidence favouring this. Han and Nambu had proposed their own quark model with integrally charged quarks [73], in contrast with the Gell-Mann–Zweig quarks which were fractionally charged. To explain the rarity of events like $K_L \rightarrow \mu^+ \mu^-$ in which strangeness quantum number changed without an accompanying change in charge, and also to achieve lepton-hadron symmetry, Glashow, Iliopoulos and Maiani [72] conjectured a fourth quark which they called *Charm*, many of whose properties they could predict. The amplitude for processes like $K_L \rightarrow \mu^+ \mu^-$ with only u, d, s quarks would be very large, but with the additional charm quark, there would be *destructive interference*. This is just the double-slit interference manifesting in a totally different context. In fact, the mass of the charm quark needed to achieve the level of suppression demanded by data at that time was estimated to be around 3 GeV. Charmed mesons as $c\bar{c}$ bound states were discovered in 1974 with a mass of 3.1 GeV by three groups, the Brookhaven Group led by Ting [74], Stanford group led by Richter [75] and a third group at Frascati [76]. The particles were called J/Ψ , and the three discovery papers appeared back to back in the same issue of Physical Review Letters. With this fourth quark, an additional bonus accrued, namely, the quarks could be fitted into the electroweak model as two doublets (u, d), (s, c). The implied mass of the charm quark was about 1.55 GeV, not too far from the estimates of Glashow–Iliopoulos and Maiani. The important lesson was to put behind SU(3) as the flavour group, notwithstanding the fact that it had such a major influence on the development of strong interaction physics. This proliferation of quarks did not continue indefinitely (or so it seems). In 1977,

Leon Lederman's team discovered the *bottom* quark [77] with a mass of 4.18 GeV, temporarily leading to an asymmetry between leptons and hadrons, which was again restored in 1995 with the discovery of the *top* quark by the CDF [78] and D0 [79] collaborations. The mass of the top was observed to be 173 GeV. This is the current fermion particle content of the so-called standard model.

The masses of the gauge vector bosons except photons were to be generated through the Higgs effect. This still left a scalar particle, called the *Higgs Boson* in the spectrum. The final triumph for this line of thinking was the experimental discovery of the Higgs particle in 2012 by the ATLAS [80] and CMS [81] collaborations. Its mass was subsequently determined to be 125 GeV.

We end this narrative with some important remarks. When the masses of the vector bosons were generated by the Higgs mechanism, the equations of motion of the theory maintained its gauge covariance. This had given rise to the hope that such massive gauge theories are actually renormalizable. An important first result in this direction was provided by Fradkin and Tyutin [82], who proved in 1969 that massless Yang–Mills theories are renormalizable. The crowning glory came in 1972 when t'Hooft and Veltman [83] showed that massive Yang–Mills theories are renormalizable too, when masses are generated by the Higgs effect. All the fermions in the standard model, like quarks and leptons too, had to get their masses through Yukawa-like couplings to the Higgs field. However, a rather technical subtlety came in the way of their proof. Due to a field-theoretic effect, called *Anomalies*, the t'Hooft–Veltman proof would be valid only if the anomalies were absent. The amazing resolution of this difficulty happens when all the three lepton and quark generations are used, and furthermore, the quarks come as colour triplets. Joglekar [84], and Cornwall, Tiktopoulos and Levin [85, 86], showed that tree-level unitarity of general theories (a much weaker criterion than renormalizability) would only be possible in the so-called *spontaneously broken gauge theories*.

It is curious that a consistent, renormalizable theory of only the weak interactions could never be found, and only unification with electromagnetism could deliver that. Despite these progresses in understanding the masses of quarks, leptons and gauge-bosons, masses of the known particles like protons, neutrons and pions could not be understood this way. Considering that the mass of a major part of the visible universe is made up of protons and neutrons, that is indeed a dampner. The rest of the book will address that issue.

References

1. H. Becquerel, Comptes Rendus **122**, 420 (1896)
2. M.P. Curie, Comptes Rendus **126**, 1101 (1897)
3. P. Curie, M.P. Curie, Comptes Rendus **127**, 175 (1898)
4. P. Curie, M.P. Curie, G. Bemont, Comptes Rendus **127**, 1215–18 (1898)
5. E. Rutherford, Phil. Mag. **47**, 109 (1899)
6. P. Villard, Comptes Rendus **130**, 1010–1012, 1178–1179 (1900)
7. H. Becquerel, Comptes Rendus **130**, 1154–57 (1900)
8. E. Rutherford, F. Soddy, Phil. Mag. **5**, 576 (1903)

9. H. Geiger, E. Marsden, Proc. Roy. Soc. Lond. A **82**, 495 (1908)
10. E. Rutherford, Lond. Edinb. Dublin Philos. Mag. Ser. 6 **21**(125), 669 (1911)
11. E. Rutherford, Proc. Roy. Soc. A **97**(686), 374 (1920)
12. W. Heisenberg, Z. f. Phy **77**, 1 (1932)
13. D. Iwanenko, Nature **129**, 798 (1932)
14. J. Chadwick, The existence of neutron. Proc. Roy. Soc. A **136**(830), 692 (1932)
15. W. Kaufmann, Physikalische Zeitschrift **4**, 54–57 (1902)
16. O. Hahn, L. Meitner, Z. f. Phy. **2**, 60 (1920)
17. L. Meitner, Z. f. Phy. **9**, 131–144 (1922)
18. L. Meitner, Z. f. Phy. **9**, 145–152 (1922)
19. J.K. Danysz, Ann. Chim. Phys. **30**, 241 (1913); Le Radium **9**, 1 (1912); Le Radium **10**, 4 (1913)
20. J. Chadwick, C.D. Ellis, Proc. Camb. Phil. Soc. **21**, 274–280 (1922)
21. C.D. Ellis, W.A. Wooster, Proc. Roy. Soc. Lond. A **117**, 109–123 (1927)
22. C.D. Ellis, W.A. Wooster, Nature **119**, 563–64 (1927)
23. C.D. Ellis, W.A. Wooster, Proc. Roy. Soc. A **114**, 276 (1927)
24. C.D. Ellis, W.A. Wooster, Proc. Camb. Phil. Soc. **23**, 717 (1927)
25. B. Swirles, Proc. Roy. Soc. A **116**, 491 (1927)
26. H.B.G. Casimir, Nature **126**, 953 (1930)
27. H.R. Hulme, Nature **126**, 643 (1930)
28. H.M. Taylor, N.F. Mott, Proc. Roy. Soc. A **138**(836), 665 (1932)
29. A. Franklin, The Spectrum of beta-decay, continuous or discrete? a variety of errors in experimental investigations, in *Boston Studies in Philosophy of Science*, ed. by G. Hon (Springer Science 2009), p. 267
30. W. Pauli, Letter **4** (1930)
31. E. Fermi, Z. f. Phy. **88**, 161 (1934)
32. L.D. Landau, Nuc. Phy. **3**, 127 (1957)
33. B. Pontecorvo, Sov. Phys. JETP **10**, 1236 (1960)
34. E.C.G. Sudarshan, R.E. Marshak, Phys. Rev. **109**, 1860 (1958)
35. R.P. Feynman, M. Gell-Mann, Phys. Rev. **109**, 193 (1958)
36. F. Reines, C. Cowan, **124**, 3212 (1956)
37. G. Gamow, E. Teller, Phys. Rev. **49**, 895 (1936)
38. T.D. Lee, C.N. Yang, Phys. Rev. **104**, 254 (1956)
39. C.S. Wu, Phys. Rev. **105**, 1413 (1957)
40. R.L. Garwin, L. Ledermann, M. Weinrich, Phys. Rev. **105**, 1415 (1957)
41. S.H. Neddermeyer, C.D. Anderson, Phys. Rev. **51**, 884 (1937)
42. H. Yukawa, Prog. Theor. Phy **17**, 48 (1935)
43. D.M. Bose, B. Chaudhuri, Nature **147**, 240 (1941); B. Chowdhury, Nature **148**, 259 (1941); Nature **149**, 302 (1942)
44. E.J. Williams, G.E. Roberts, Nature **145**, 102 (1940)
45. F. Rasetti, Phys. Rev. **59**, 613 (1941); Phys. Rev. **60**, 198 (1941)
46. M. Conversi, E. Pancini, O. Piccini, Nuovo Cimento **3**, 372 (1946); Phys. Rev. **71**, 209 (1947)
47. J. Steinberger, Phys. Rev. **74**, 500 (1948)
48. C.M.G. Lattes, H. Muirhead, G.P.S. Occhialini, C.F. Powell, Nature **159**, 694–697 (1947)
49. R. Bjorklund, W.E. Randall, B.J. Moyer, H.F. York, Phys. Rev. **77**, 213 (1950)
50. J. Schwinger, Ann. Phys. NY **2**, 407 (1957)
51. S. Lokanathan, J. Steinberger, Phys. Rev. **98**, 240 (1955)
52. G. Feinberg, Phys. Rev. **110**, 1482 (1958)
53. G. Danby, J.-M. Gaillard, K. Goulianos, L.M. Lederman, N. Mistry, M. Schwarz, J. Steinberger, Phys. Rev. Lett. **9**, 36 (1962)
54. M.L. Perl et al., Phys. Rev. Lett. **35**, 1489 (1975)
55. O. Klein, in *New Theories in Physics*, Warsaw, 30 May–3 June (1938)
56. C.N. Yang, R. Mills, Phys. Rev. **95**, 631 (1954); Phys. Rev. **96**, 191 (1954)
57. J. Reiff, M. Veltman, Nuc. Phy. B **13**, 545 (1969)
58. W. Heisenberg, Z. f. Phy. **49**, 619 (1928)

59. Y. Nambu, Phys. Rev. **117**(3), 648 (1960)
60. J. Goldstone, Nuovo Cimento **19**(1), 154 (1961)
61. F. Englert, R. Brout, Phys. Rev. Lett. **13**, 321 (1964)
62. P.W. Higgs, Phys. Rev. Lett. **13**, 508 (1964)
63. G.S. Guralnik, C.R. Hagen, T.W.B. Kibble, Phys. Rev. Lett. **13**, 585 (1964)
64. S. Bludman, Nuovo Cimento **9**, 433 (1958)
65. S.L. Glashow, Nuc. Phys. **10**, 107 (1959)
66. A. Salam, J.C. Ward, Nuovo Cimento **11**, 568 (1959)
67. S. Weinberg, Phys. Rev. Lett. **19**, 1264–66 (1967)
68. M. Gell-Mann, Phys. Letts. B **8**, 214 (1964)
69. G. Zweig, CERN Reports 8182/TH401, 8419/TH412
70. H. Fritzsch, M. Gell-Mann, H. Leutwyler, Phys. Lett. B **47**, 4 (1973)
71. J. Bjorken, S. Glashow, Phys. Lett. **11**, 255 (1964)
72. S.L. Glashow, J. Iliopoulos, L. Maiani, Phys. Rev. D **2**, 1285 (1970)
73. M.Y. Han, Y. Nambu, Phys. Rev. **139**, B1006 (1965)
74. J.J. Aubert et al., Phys. Rev. Lett. **33**, 1404 (1974)
75. J.E. Augustin et al., Phys. Rev. Lett. **33**, 1406 (1974)
76. F. Collaboration, Phys. Rev. Lett. **33**, 1408 (1974)
77. S.W. Herb et al., Phys. Rev. Lett. **39**, 252 (1977)
78. F. Abe et al., Phys. Rev. Lett. **74**, 2626 (1995)
79. S. Abachi et al., Phys. Rev. Lett. **74**, 2632 (1995)
80. ATLAS Collaboration, G. Aad et al., Phys. Lett. B **716**, 29 (2012)
81. CMS Collaboration, S. Chatrchyan et al., Phys. Lett. B **716**, 31 (2012)
82. E.S. Fradkin, I.V. Tyutin, Phys. Lett. B **30**, 562 (1969)
83. G. t'Hooft, M.T. Veltman, Nuc. Phys. B **44**, 189 (1972)
84. S.D. Joglekar, Ann. Phys. **83**, 427 (1974)
85. J.M. Cornwall, G. Tiktopoulos, D.N. Levin, Phys. Rev. Lett. **30**, 1268 (1973); Phys. Rev. Lett. **31**, 572(E) (1973)
86. J.M. Cornwall, G. Tiktopoulos, D.N. Levin, Phys. Rev. D **10**, 1145 (1974); Phys. Rev. D **11**, 972(E) (1975)

Nuclear Forces, Meson Field Theories and Their Failures

3

3.1 Nuclear Forces: Observational

With the recognition by Heisenberg [1] and Iwanenko [2] that nuclei are made up of protons and neutrons, the natural next step was in understanding how the protons and neutrons are held together in a nucleus. Also, the question of what offsets the coulomb repulsion between protons (rather sizeable at the short separations involved) was part of that. It is useful to recount several important observational inferences just prior to Yukawa's conceptual breakthrough in 1935 [3]. The reader is referred to the exhaustive account by Bethe and Bacher [4], called *Bethe's Bible* in a lighter vein, as also Mukherji's historical commentary [5], which is also technically faithful.

- The nuclear forces are strong. The very fact that a very compact electrically charged nucleus does not fly apart due to coulomb repulsion shows that the nuclear forces must be rather strong, and attractive.
- The nuclear forces are *saturating*. The most straightforward characterization of saturation is that the average binding energy per particle approaches a constant value, i.e. it saturates. The binding energy per nucleon in deuteron is about 1.1 Mev. This steadily increases, though not *monotonically*, as the atomic mass A increases, and reaches a value of about 7–8 MeV per nucleon.
- Saturation is only possible with short-range forces. In contrast, the electrostatic energy between Z particles of charge e grows as $Z(Z-1)$ and the average energy per particle *grows* as Z , for large Z . Same holds for gravitational interactions.
- Empirically, the evidence for saturation and short-range of nuclear forces came from the fact that binding energies grew linearly with A . In the famous *Empirical Mass Formula* of Weizsäcker [6], the leading term is proportional to A . The precursor to this was Gamow's work [7] based on the so-called liquid-drop model of the nucleus. Of course, there are many sub-leading terms involving the atomic number Z also, but we shall not go into those details.

- The inferred range of nuclear forces was about $2 \cdot 10^{-13}$ cm. This was inferred mostly from nuclear scattering data. The earliest scattering experiment was with protons and protons, in 1936, by Tuve and Heydenburg [8]. These clearly indicated the presence of an attractive, shortrange force in addition to the repulsive and long-range coulomb potential.
- Yukawa in his 1935 seminal work [3] on meson exchange theory of nuclear force had used this value for the range which would roughly translate into a mass 200 times the electron mass or, about 100 MeV.
- Nuclear forces have a spin dependence. An indication of this was the fact that the deuteron ground state was in the *Triplet Spin State* where the proton and neutron spins are parallel. If the nuclear forces were totally spin-independent, one would expect both the triplet state and the singlet state in a ratio of 3:1. Already in 1932, Heisenberg [1] had proposed the nuclear forces to be spin-dependent, while Majorana in 1933 had proposed spin-independent nuclear forces [9, 10].
- One of the most important inferred properties of nuclear force was that of *charge independence*. Roughly stated, it implied that neutron-neutron, proton-proton and neutron-proton forces are comparable. This of course requires ignoring the coulomb forces between protons. A good test of this would be provided by the difference in binding energies of Tritium (H^3) and He^3 . If nuclear forces were indeed charge-independent, this difference should be entirely attributable to coulomb energy of the proton pair in He^3 (and of course the observed mass difference between protons and neutrons). Such a calculation was indeed performed and found to support the charge independence (see Sect. 22 of Bethe-Bacher [4]). Kemmer had provided the theoretical formalism for charge independence in 1937 and a reformulation in the light of Yukawa's mesons [11].
- A quantitative measure of the strength of the nuclear forces was also provided by the observed meson-nucleon cross sections of the order of 10^{-28} cm².

Even before Yukawa, many proposals had been made to explain the forces between nucleons that would keep them bound within the small nucleus. One of the earliest was the *electron-exchange* model of Heisenberg, put forward in 1932 [1]. According to this, a neutron would emit an electron to become a proton, and the electron would be absorbed by a proton to become a neutron. This model suffered from several deficiencies. The magnitude of the force comes out too small. Even though Pauli [12] had already proposed the neutrino to overcome the spin and statistics difficulties of beta-decay, that idea had not gained much currency as can be seen from the fact that even Heisenberg had not invoked the neutrinos in this model. Consequently, this model violates angular momentum conservation. Since the exchanged particle is electron, it also predicts a range some 200 times *larger* than observed. The observed saturation of nuclear forces would also not set in for the range of A experimentally observed.

Few years later, in 1934 and after Fermi had proposed his theory of beta-decays [13], Heisenberg himself improved this model by making the (e, ν) field (the Fermi field) as the carrier of nuclear forces. The angular momentum and statistics difficulties of his 1932 model disappeared. Iwanenko [14], Nordsieck [15] and Tamm [16]

analysed Heisenberg's proposal in detail the same year, and found many problems with it. Firstly, the magnitude of the nuclear force was about 10^{12} times too small; this was due to the very small value of the coupling in beta-decays. Equally seriously, the nuclear forces diverged as $r \rightarrow 0$.

3.2 Yukawa Theory

In 1935, Yukawa proposed his meson theory of nuclear forces [3]. The principal idea was that the range of a force is inversely proportional to the mass of the carrier, i.e. $a = \frac{\hbar}{mc}$. Based on the accepted value of the range, Yukawa chose the mass for his carriers to be around 200 times the electron mass. He chose the simplest possibility for the carrier field, namely, a relativistic scalar field with mass. The spin of the carrier was consequently taken to be zero. It obeyed the Klein-Gordon equation named after Klein [18] and Gordon [19] who had proposed it as a relativistic description of electrons. It may be recalled that Schrödinger in 1925, a year before Klein and Gordon, had first chosen the same equation to base his quantum theory of the hydrogen atom, and had abandoned it as it failed to account for the observed fine structure.

Though Yukawa treated the carriers relativistically, he treated the nucleons non-relativistically. So the resulting nuclear forces did not have any spin dependence. The pions, the name for the Yukawa mesons, were charged. Yukawa only considered the neutron-proton interactions. Subsequently, in 1937, Yukawa treated the nucleons also relativistically while maintaining the scalar nature of pions [20]. The resulting spin dependence of the nuclear forces was incorrect, predicting repulsive forces for the triplet states of deuteron, and attractive forces for the singlet state.

Just a year earlier, in 1936, Proca [21] had argued that there were other spin possibilities for the carriers, and in particular, a massive vector field would do too. We have already mentioned how Oskar Klein had chosen such a possibility in 1938 [22] for mediating the beta-interactions! Fröhlich [23] analysed the vector hypothesis for its consequences, and found that, while they predicted attractive forces for the triplet states of deuteron, as desired, they also predicted attraction for the singlet state too. Bhabha was also one of the proponents of the vector theories [24]. But for non-relativistic nucleons, the vector and scalar (but not pseudo-scalar) are practically indistinguishable. Kemmer, in 1938, went even further and proposed that scalar (S), pseudo-scalar (P), vector (V) and axial-vector (A) were all possibilities [25]. After analysing the phenomenological implications, Kemmer favoured the vector nature. Though he was the first to seriously consider the pseudo-scalar option, the one that triumphed in the end, Kemmer kept favouring the vector option, as it seemed to provide a natural explanation for the magnetic moments. The thought in those days was that the carriers themselves should possess magnetic moments in order to account for the anomalous values. We know today that that is not really correct and even scalar or pseudo-scalar carriers can contribute to the magnetic moments as long as they are charged. According to Mukherji [5], Pauli was convinced of the advantages of the pseudo-scalar theory; according to Pauli, the n-p interactions with a triplet of pions

would only leave the pseudo-scalar option. In 1941, Robert Oppenheimer argued that only pseudo-scalar pions could explain the correct spin dependence of nuclear forces [26].

As Yukawa had not treated the “like-particle” interactions like, p-p and n-n, he either failed to consider the implications of charge independence of nuclear forces or he was unaware of it at that time. With only charged pions, the n-n and p-p interactions were not possible in the first order. However, at the second order, even the charged Yukawa-mesons could accommodate them. But they would be of order g^4 in Yukawa-coupling unlike n-p which was of order g^2 , so charge independence would still be difficult to explain on the basis of the Yukawa theory. Furthermore, calculations showed that the resultant like-particle interactions were repulsive, like the electromagnetic interactions. In 1938, Tuve, Hafstad and Heydenburg [27] repeated their earlier experiments on p-p scattering and showed clearly that the short-range part of p-p interactions was attractive. To resolve these difficulties, Kemmer, in 1938, introduced a neutral pion into the Yukawa scheme [28]. Even more remarkably, he suggested that the three mesons together are part of an *isotopic spin triplet*. Recall that it was Heisenberg who had introduced the isotopic spin concept, and had treated neutrons and protons as components of an isotopic spin doublet. Shortly afterwards, Yukawa, Sakata, Kobayasi and Taketani also introduced the neutral pion [29]. An interesting postscript was added by Julian Schwinger in 1950 to the issue of charge independence of nuclear forces [30]. By that time, there was a small but noticeable deviation from strict charge independence. Schwinger showed that the deviations can be well accounted for by magnetic interactions due to the anomalous magnetic moments of nucleons. Consensus in favour of isospin triplet of pseudo-scalar pions was slowly building up.

3.3 Meson Field Theories and Their Failures

Amidst all these discussions about the nature of Yukawa’s mesons, efforts were also being made towards field theories of nuclear forces. It should be remembered that around this time even the quantum field theories of electromagnetism were hardly in a satisfactory stage. But it was already a decade since the pioneering efforts of Dirac, Fock and others [31–33], as already recounted before. The experimental discovery of muons in 1936 with markedly different features from what was anticipated of Yukawa mesons further confounded the issues. It was not clear which of these mesons (μ or π) mediated the nuclear forces. A variety of so-called “two-meson theories” made the situation very confusing. Among the pions too, there was no clarity if both charged and neutral pions participated. Even possibilities of only neutral pions being the carriers were seriously considered, as also “mixed meson theories” where carriers of various spins were considered simultaneously.

The field theories were expected to provide quantitatively satisfactory accounts of such observed phenomena as the deuteron ground state and its observable properties, the magnetic and quadrupole moments of nucleons, meson-nucleon and nucleon-nucleon scattering, the strength and nature of the meson-nucleon interaction, etc.

One of the earliest attempts at formulating a quantum field theory of nuclear forces was undertaken by Lamb and Schiff in 1938 [34]. They had set up a formalism for scalar carriers that was very close in spirit to Dirac's pioneering ideas on quantizing electromagnetic fields in 1927 [32]. Their aim was to account for the magnetic moments of nucleons, and as they erroneously argued that scalar-charged mesons can not contribute, they were not very positive about the Yukawa theory. Instead, they considered six component vector carriers in addition to the scalars. Though most of the details of this work did not survive posterity, some of their motivations were right on target!

They expressed the hopes of finding a field theory that would account for both the observed features of nuclear forces(interaction), as well as such properties as magnetic moments, quadrupole moments(structure), etc. They expected such a field theory to be convergent, and account for strong, short-range forces. In addition, they express the desirability of such a field theoretic description to be covariant and gauge-invariant.

Bethe in 1938 (going on till 1940) based his field theories only on neutral mesons, which he assumed to be spin-1. His 1940 works [35] found some successes even though the meson fields had been treated classically. However, the tensor interactions diverged at the origin, and cut-offs were necessary. In his 1939 work [36], even though he found satisfactory values for the deuteron quadrupole moment, he found disagreements with beta-decay as well as the nucleon magnetic moments.

Moeller and Rosenfeld in 1939 [37] proposed the mixed-meson theories with vector and pseudo-scalar mesons. But their theory was also inadequate to explain all the observed features. They ascribed quadrupole moments to relativistic effects. There were many variants of field theories such as those favoured by Taketani [38]. Meson-pair theories were proposed to obviate the difficulties of not having neutral mesons.

Models were also proposed where the coupling could be weak, intermediate or strong. Some people were of the opinion that the meson-nucleon problem was much harder than the corresponding quantum electrodynamics problem because of the "finite extension" of the nucleons versus the point electron. This is really a moot point as it is not clear if the observed finite extension is a consequence of a point bare nucleon interacting very strongly with mesons.

As late as 1946, Kusaka [39] concluded that no satisfactory theory of nuclear forces was possible based on conventional meson field theories. This was despite the best efforts by such eminent physicists as Yukawa, Serber, Sakata, Teller, Bethe, Oppenheimer, Majorana, Heisenberg, Bhabha, Pauli, etc.

3.4 Experimental Discoveries of the Pions

Charged pions were discovered by Lattes, Occhialini, Muirhead and Powell in 1947 [40]. Interestingly, the discovery came immediately after the famous Shelter Island meeting, which we have discussed in the context of the development of Quantum Electrodynamics. The observed mass was 139.570 MeV or $279 m_e$. It was strongly

absorbed by matter compared to the cosmic meson(muon). The charged pions were observed to undergo their own beta-decay with a lifetime of about $2.6 \cdot 10^{-8}$ s, a hundred times shorter than muons. The decay products were predominantly muons and muon-neutrinos, not electrons and electron-neutrinos. This was a confirmation of the V-A structure of weak currents. The strength of the coupling was the same as in nuclear beta-decay (we are ignoring details like Cabibbo angles, etc.). So, the pion decay was essential to establishing the universal nature of weak interactions.

The neutral pions took nearly three more years to be discovered. They were discovered by Bjorklund et al. in 1950 [41]. The experiment consisted of careful determination of gamma-ray yields in collisions of protons on nuclear targets. The excess gamma rays were interpreted as the decay products of neutral pions which were photoproduced in the reaction. The mode is now understood to be decay of neutral pions into two photons. This interpretation was put forward theoretically by Primakoff [42]. The current estimate of the lifetime is about $8.4 \cdot 10^{-16}$ s. Just 2 years after Kemmer had proposed the neutral pion in 1938 [28] to explain charge independence in nuclear forces, Sakata and Tanikawa, in 1940, predicted theoretically the neutral pion decay rate to be about $\simeq 10^{-16}$ s [43]. This decay came to play a rather central role in the subsequent development of theoretical physics. For a fascinating account of this, see Miskimen [44].

The spin of the charged pions was determined by Durbin, Loar and Steinberger in 1951 [45] by studying the dissociation of deuteron upon bombardment by π^+ into two protons, and its inverse reaction. The spin was found to be zero. The spin of the neutral pion was determined by a detailed analysis of the two-photon decay mode. This also determined the neutral pion to be a pseudo-scalar as conjectured by Pauli long ago. But the latter analysis was rather detailed and had to assume that parity was not violated in electromagnetic interactions. This analysis was carried out by Plano et al. [46]. The charged pions were determined to be pseudo-scalar by Panofsky, Aamodt and Hadley in 1951 by studying the disintegration of deuterons into two neutrons by absorption of very low energy negatively charged pions [47].

3.5 Meson Theories Post-Pion Discovery

By the time the pions were discovered and their properties elucidated it was close to the mid-fifties. All the technical and conceptual problems in the development of Quantum Electrodynamics had already been overcome. One might have thought that with those breakthroughs the field theory of nuclear forces would also have moved forward. But that was not to be. The large strength of pion-nucleon interactions made the perturbative techniques so essential for the breakthroughs in QED, not very reliable. Surprisingly, even though the (pseudo)scalar interactions ought to have been simpler than gauge interactions of QED, they had many subtle difficulties of their own.

An interesting early attempt was by Geoffrey Chew and Francis Low in 1956 that came to be called the “Chew-Low theory” [48]. It combined Chew’s earlier work on his “static model” [49]. They treated the nucleons as static, but the pions as relativistic. The interaction was taken to be of the “pseudo-vector” type. It was an old-fashioned type field theory with some flavour of the S-matrix methods that were to dominate the theoretical discourses soon. It was called an “effective range” approach to low-energy p-wave scattering only. It still needed cut-offs. They claimed that many features of pion-nucleon scattering, photo production of pions, nucleon-nucleon scattering, deuteron ground state as well as the nucleon magnetic moments were well captured by this approach. Low also had applied Dyson’s perturbative methods of QED to develop an S-matrix. But the strength of the meson-nucleon coupling made the applicability of such an S-matrix rather doubtful.

Some comments on the nature of the pion-nucleon interactions are in order at this point. As long as nucleons are “on mass-shell”, meaning they obey the relativistic energy-momentum relations, the interaction can be equivalently described as either pseudo-scalar or pseudo-vector with gradient coupling to pion fields. The former has very large coupling, making perturbative approaches useless, while the latter is non-renormalizable.

In the early calculations of the two-photon decays of the neutral pion, there were disagreements between these two types of interactions as could be seen by the works of Fukuda and Miyamoto [50] on the one hand, and of Jack Steinberger [51] on the other. This was resolved by Schwinger through the application of his very elegant gauge invariant methods which he applied to the interactions being scalar, pseudo-scalar or pseudo-vector [52]. In particular, he showed that the pseudo-scalar and pseudo-vector interactions give the same answer in the leading approximation. The results of this paper were also very important in the determination of the parity of the neutral pion.

Already by 1947, Rochester and Butler [53] had discovered a whole new class of mesons that were subsequently named “strange mesons” because they could be produced as pairs via strong interactions but only decay through weak interactions. While they would be unimportant for meson field theories at the level of one-meson exchanges, any relativistic field theory had to account for their associated production. This further muddled the prospects for a relativistic quantum field theory of nuclear forces.

Another development was the entry of Rho-mesons into the picture. Conjectured by Nambu in 1957 [54] to account for the charge distributions of nucleons, it was experimentally discovered by Erwin, March, Walker and West in 1960 [55]. It had a mass of 770 MeV, nearly five times the mass of pions. There was no more justification to build a strong-interaction theory based only on pions and nucleons.

Attention soon shifted to analytic S-matrix techniques [56]. A pioneering work was by Gell-Mann, Goldberger and Thirring [57] leading to the concept of Dispersion Relations (an exhaustive treatment of this will be given in Chap. 11). An early application of Dispersion Relations to pion-nucleon scattering by Chew, Goldberger, Low and Nambu [58] started a new era in strong-interaction physics. That will be the focus of the main part of this book, so also another radical approach pioneered

by Gell-Mann and Zweig [59,60] by way of their quark model which culminated in a relativistic quantum field theory of strong interactions called Quantum Chromodynamics.

References

1. Werner Heisenberg, Z.f. Phy. **77**, 1 (1932)
2. D. Iwanenko, Nature **129**, 798 (1932)
3. H. Yukawa, Prog. Theor. Phy. **17**, 48 (1935)
4. H.A. Bethe, R. Bacher, Rev. Mod. Phys. **8**, 32 (1936)
5. V. Mukherji, A History of the meson theory of nuclear forces from 1935 to 1952. Arch. Hist. Exact. Sci. **14**, 27–102 (1974)
6. C.F. Weizsäcker, Phys. Zeits. **36**, 779 (1935); Z. f. Phy. **96**, 431(1935)
7. G. Gamow, Proc. Roy. Soc. A **123**, 386 (1929); Proc. Roy. Soc. A **126**, 803 (1930)
8. H. Tuve, Heydenburg, Phys. Rev. **49**, 432 (1936); Phys. Rev. **50**, 806 (1936)
9. E. Majorana, Z. f. Phys. **82**, 137 (1933)
10. E. Majorana, Kerntheorie. Z. f. Phys **82**, 137 (1933)
11. N. Kemmer, Phys. Rev. **55**, 105 (1937)
12. W. Pauli, Letter (1930). Accessed from 4 Dec 1930
13. E. Fermi, Z. f. Phy. **88**, 161 (1934)
14. D. Iwanenko, Nature **133**, 981 (1934)
15. A.T. Nordsieck, Phys. Rev. **46**, 234 (1934)
16. I. Tamm, Nature **133**, 981 (1934)
17. E. Schrödinger, unpublished (1935)
18. O. Klein, Z. f. Phy. **37**, 895 (1926)
19. W. Gordon, Z. F. Phy. **40**, 117 (1926)
20. H. Yukawa, Prog. Theor. Phy. **19**, 1084 (1937)
21. A. Proca, J. Phys. Radium **7**, 347 (1936)
22. O. Klein, in *New Theories in Physics* (Warsaw, 1938). Accessed from 30 May-3 June 1938
23. H. Fröhlich, Kemmer, Proc. Roy. Soc. A **166**, 154 (1938)
24. H.J. Bhabha, Proc. Roy. Soc. **166**, 501 (1938); Proc. Roy. Soc. **172**, 384 (1939)
25. N. Kemmer, Proc. Roy. Soc. **A166**, 127 (1938)
26. R. Oppenheimer, Phys. Rev. **59**, 462 (1941)
27. H. Tuve, Heydenburg, Phys. Rev. **53**, 239 (1938)
28. N. Kemmer, Proc. Camb. Phil. Soc. **34**, p354 (1938)
29. H. Yukawa, S. Sakata, Kobayasi and Taketani. Prog. Theor. Phys. **20**, 720 (1938)
30. J. Schwinger, Phys. Rev. **78**, 135 (1950)
31. P.A.M. Dirac, Proc. Roy. Soc. A **112**(762), 661 (1926)
32. P.A.M. Dirac, Proc. Roy. Soc. London A **114**, 243 (1927)
33. P.A.M. Dirac, V.A. Fock, B. Podolsky, Phys. Zeits. d. Sowjetunion **2**, 468(1932); **3**, 64 (1932)
34. W. Lamb, L. Schiff, Phys. Rev. **53**, 651 (1938)
35. H.A. Bethe, Phys. Rev. **57**, 260, 390 (1940)
36. H.A. Bethe, Phys. Rev. **55**, 1261 (1939)
37. C. Moeller, L. Rosenfeld, Nature **143**, 241 (1939); Nature **144**, 476 (1939)
38. M. Taketani, Prog. Theor. Phy. **7**, 45 (1952)
39. A. Kusaka, Phys. Rev. **70**, 794 (1946)
40. C. Lattes, G. Occhialini, H. Muirhead, C. Powell, Nature **159**, 694 (1947)
41. R. Bjorklund, W.E. Crandall, B.J. Moyer, H.F. York, Phys. Rev. **77**, 213 (1950)
42. H. Primakoff, Phys. Rev. **81**, 899 (1951)
43. S. Sakata, Y. Tanikawa, Phys. Rev. **57**, 548 (1940)
44. R. Miskimen, *Ann (Rev. Nucl. Part, Sci)*, 2011)
45. R. Durbin, H. Loar, J. Steinberger, Phys. Rev. **84**, 581 (1951)

46. R. Plano, A. Prodell, N. Samios, M. Schwarz, J. Steinberger, Phys. Rev. Lett. **3**, 525 (1959)
47. W. Panofsky, W. Panofsky, R.L. Aamodt, J. Hadley, Phys. Rev. **81**, 565 (1951)
48. G.F. Chew, F.E. Low, Phys. Rev. **101**, 1570 (1956)
49. G.F. Chew, Phys. Rev. **95**, 1669 (1954)
50. H. Fukuda, Y. Miyamoto, Prog. Theor. Phys. **4**, 342 (1949)
51. J. Steinberger, Phys. Rev. **76**, 1180 (1949)
52. J. Schwinger, Phys. Rev. **82**, 664 (1951)
53. G.D. Rochester, C.C. Butler, Nature **160**, 855 (1947)
54. Y. Nambu, Phys. Rev. **106**, 1366 (1957)
55. A.P. Erwin, R. March, W.D. Walker, E. West, Phys. Rev. Lett. **4**, 142 (1960)
56. F.E. Low, Phys. Rev. **97**, 1392 (1955)
57. M. Gell-Mann, M.L. Goldberger, W. Thirring, Phys. Rev. **95**, 1612 (1954)
58. G.F. Chew, M.L. Goldberger, F.E. Low, Y. Nambu, Phys. Rev. **106**, 1337 (1957)
59. M. Gell-Mann, Phys. Lett. B **8**, 214 (1964)
60. G. Zweig, CERN Reports 8182/TH401 1964

Part II

Heisenberg's S-matrix to String Theory



The S-matrix: From Heisenberg Till Now

4

4.1 Introduction

The concept of an S-matrix in nuclear physics was first discussed by Wheeler [1]. This was essentially an extension of the idea of transition matrix elements in quantum mechanics. One very important difference between quantum mechanics and quantum field theory is that in QM, the initial and final states can be any arbitrary pair of quantum states, while in field theory they have to be the so-called *free asymptotic states*. The classical analog is somewhat elementary, as a given initial state evolves to another definite final state. So, the S-matrix has exactly one entry in each row(column); this can also be seen as the reflection of the deterministic nature of classical theory. We will be mostly concerned with the S-matrix in quantum field theories. It was Heisenberg who introduced the S-matrix in field theories in 1942. In English, S-matrix has the meaning of a *Scattering Matrix*, in the sense that it codifies all the scattering information of a theory. In German, it is called *Streu-Matrix*, an S-matrix there too!

Heisenberg's pursuit of the S-matrix was rooted in his thinking about the divergence difficulties of quantum field theories. He felt (though modern thinking on this has been very different) that the divergence difficulties were an indication of a *fundamental length* in physics. More specifically, he felt that divergences arose from arbitrarily short distances (he was in sync with the most modern views here). Heisenberg did not think that interactions were that crucial in this context. He thought that the fundamental scale was $\simeq 10^{-13}$ cms, a scale well surpassed in high energy physics without encountering any fundamental lengths. This does not necessarily negate the idea of a fundamental length scale. In fact, there are some who think that the so-called *Planck Scale* of roughly 10^{-33} cms is indeed a fundamental length. Heisenberg took this idea quite seriously and constructed field-theoretic models (as part of his unified field theory project) with explicit appearance of a fundamental length scale. What he did, in essence, was admitting what would be described today

as *non-renormalizable* interactions. It is an intriguing thought whether the methodology he suggested to tackle this, namely, S-matrix theory, would be right to understand non-renormalizable field theories!

In this context, Heisenberg raised some questions which are very much relevant even to this day. In particular, "...which concepts of existing theory of wavefields (read Quantum Field Theory) would survive in a more comprehensive theory". He considered as important what he called the *Observable Quantities*, and listed energy eigenvalues of closed systems, probabilities for collisions, absorption, emission, etc. These are associated with the *asymptotic behaviour* of wavefields and are characterized by a *matrix* labelled by energy-momentum, spin and other quantum numbers. The importance of asymptotic fields lies in the fact that for them the existence of such conserved quantities is easier to establish. Heisenberg viewed this matrix, what came to be called the S-matrix, as a primary quantity to be calculated directly.

A good source on Heisenberg and the S-matrix can be found in Oehme's historical account [2]. Unfortunately, it lacks important references. Heisenberg's first published account of the S-matrix is in his 1942 (published in 1943) paper in *Zeitschrift fur Physik* [3]. He later followed it up with a number of papers and lectures. His 1942 paper was entitled *Observable Quantities in the theory of elementary particles*. In this work, he explored the implications of (i) Poincare invariance (it should be carefully noted that this in itself does not require a space-time description) and (ii) conservation of probability equivalently the requirement of *Unitarity*. He mathematically formulated the latter as

$$S = e^{i\eta} \quad (4.1)$$

One of Heisenberg's most important quests was η . Despite its trivial looks, Eq. (4.1) is very difficult as it is an operator equation operating on a (Hilbert)-space infinitely more complex than what one encounters in Quantum Mechanics! In his second paper [4], Heisenberg considered some simple *Ansätze* for η and worked out their consequences for simple scattering and production processes. C. Moeller had also worked out the connections between the S-matrix and such observable processes [5,6]. According to Oehme, H.A. Kramers informed Heisenberg of the *analytical properties* of the S-matrix as a function of the momentum variables (treated as complex) with the physical amplitudes as a function of real momentum values arising out of the restriction of the analytic functions to the real axis. Kramers is said to have obtained these results as an extension of his work with S.A. Wouthuysen in 1940. Kramers is also supposed to have told Heisenberg that the simple poles of the S-matrix are connected with the single-particle states of the theory.

In the light of this, Heisenberg, in his third paper [7] expanded the scope of his S-matrix program to include analyticity properties. This was indeed a gamechanger. Nevertheless, there were many criticisms of his S-matrix theory. Res Jost seemed to have found "false poles", and unwanted zeroes of the S-matrix, as also Pauli and Ma [8]. These difficulties prompted Pauli to reject Heisenberg's ideas as a failure. Oehme claims that Heisenberg was not particularly worried about these false zeroes and singularities, and had instead expressed hopes that future developments would clarify these issues. Indeed, the false singularities were later identified with the

physical singularities of the crossed channels. As we shall soon see, the so-called *crossing symmetry* is another crucial ingredient in the construction of the S-matrix, along with unitarity, Lorentz-invariance and analyticity.

In spite of all these developments, it was not clear how the S-matrix was to be determined in practice. The situation was most succinctly summed up by the pointed criticism of Pauli “....but they remain only a program as long as no method is given for a theoretical determination of the S-matrix” [2]. After the war also Heisenberg continued to be preoccupied with the S-matrix. He published several important papers in the process. Among them are his paper in 1946 on the mathematical framework of the quantum theory of wavefields (read quantum field theory) [9], his lecture on “The Present Situation in the Theory of Elementary Particles” at Cambridge in 1947 [10], and his Reviews of Modern Physics article in 1957 titled “Quantum Theory of Fields and Elementary Particles” [11]. Another interesting work around that time was by Eden [12].

In the aforementioned works, Heisenberg made a number of deep remarks about field theories and elementary particles. He sought to draw a distinction between divergences occurring in meson field theories and those occurring in Quantum Electrodynamics. As examples of the “non-uniqueness” of the quantization procedure, he cites the Dirac and Klein-Gordon equations. Rather surprisingly, he says that even the reverse process of recovering the classical limit from a quantum theory is not unique. He explicitly states that “For one set of quantum theoretic equations, one may find several classical ‘pictures’”. As an example, he says the quantum mechanical state of an atom, in the classical limit, can either be pictured as an electron moving around a nucleus, or, as standing waves of electrons. Indeed, if wave-particle duality is a manifestation of a quantum theory, this has to be so. Among other things, he tries to give an intuitive picture of divergences in quantum field theories. He also raises questions about the distinction between elementary and composite systems in relativistic field theories. For example, is a neutron an elementary particle, or a bound state of a proton and a negatively charged pion? This is very much in the spirit of what Geoffrey Chew called *Nuclear Democracy* as a fundamental distinction between S-matrix theories on the one hand and Quantum Field Theories of Elementary particles on the other.

4.2 Kramers, Kronig and Analyticity

The next dramatic developments in S-matrix theory were due to a suggestion by Kronig [13] to impose further restrictions on the S-matrix of the type seen in optics called *Dispersion Relations*. Though he made this in a very short ‘Letter to the Editor’, its impact has been long-lasting. These are relations between the real and imaginary parts of a complex function. In optics, this relation is between the real and imaginary parts of the dielectric ‘constant’ $\epsilon(\omega)$. There, the real part is the refractive index, while the imaginary part is absorption. That refractive index and absorption can be clubbed into a single complex function follows from their very definitions.

Kronig's specific suggestion was that Heisenberg's S-matrix elements should satisfy similar dispersion relations.

It is appropriate at this point to draw a clear distinction between *Dispersion* and *Dispersion Relations*. The former refers to the fact that in general refractive index and absorption coefficient are frequency-dependent. The latter refers to a relationship between the imaginary and real parts of $\epsilon(\omega)$, and as we shall see shortly, is considerably deeper conceptually. Mathematically, such relations are called *Hilbert Transforms*, and they are consequences of certain analyticity properties when ω is extended to *complex* values. This analyticity is itself a consequence of the physical principle of *causality*. Therefore, Kronig's suggestion was tantamount to using causality, and the resulting analyticity in constraining the S-matrix.

Dispersion was investigated as early as the 1900s by Drude and Lorentz, and in classical electrodynamics typically takes the form (see for example, Eq. (7.51) [14])

$$\epsilon(\omega) = 1 + \frac{4\pi N e^2}{m} \sum_i \frac{f_i}{\omega_i^2 - \omega^2 - i\omega\gamma_i} \quad (4.2)$$

where N is the number of atoms (molecules) per unit volume, and the electrons are bound by damped harmonic oscillators of frequencies ω_i and damping γ_i . The f_i are called *oscillator strengths* obeying

$$\sum_i f_i = Z \quad (4.3)$$

with Z the number of electrons per atom/molecule. During the development of Quantum Mechanics, Kramers and Heisenberg [15] had proposed the quantum version of this dispersion formula which was very influential.

Dispersion relations on the other hand are totally different. In this classical electrodynamical context, they take the form (see Eq. (7.120) of [14])

$$\begin{aligned} \text{Re } \epsilon(\omega) &= \frac{2}{\pi} P \int_0^\infty \frac{\omega' \text{Im } \epsilon(\omega')}{\omega'^2 - \omega^2} d\omega' \\ \text{Im } \epsilon(\omega) &= -\frac{2\omega}{\pi} P \int_0^\infty \frac{\omega' \text{Re } \epsilon(\omega')}{\omega'^2 - \omega^2} d\omega' \end{aligned} \quad (4.4)$$

Kronig was among the first to state this relation between the real and imaginary parts in his 1926 paper [16] on the dispersion of X-rays. The said relation between real and imaginary parts appears in Eq. (9) of the paper. Ladenberg had, as early as 1921, stated this *intimate connection* between dispersion (of the refractive index) and absorption [17, 18]. Even though Kronig's paper had discussed the relationship between refractive index and absorption, it did not explicitly bring out the analyticity and causality aspects. Kramers's inputs to Heisenberg about the importance of the analytic properties of the S-matrix have already been alluded to earlier. His work that had a central and decisive impact on the S-matrix program was his paper in 1927 titled *La diffusion de la lumiere par les atomes* [19]. Unfortunately for those wishing

to access such original sources, this highly influential work was part of a relatively unknown (but very important) conference in Como, Italy.¹ Though Kramers was a Dutchman, he published this pathbreaking work of his in the proceedings of an Italian conference and the language of his paper was French!

Apart from clear exposition of dispersion, both classically and quantum mechanically (his work with Heisenberg [15]), he goes on to show the explicit dispersion relations, that is, the relationship between the real and imaginary parts of $\epsilon(\omega)$ in his Eq. (18). He acknowledges the work of Kronig [16], as well those of Kallmann and Mark [20], also in the context of dispersion in X-rays. Immediately afterwards, he points out that dispersion relations of this type have a very simple mathematical significance (*une signification mathématique tres simple*). This is the aspect of *analyticity*. Immediately, after his Eq. (21), Kramers states that the refractive index and absorption constitute a single complex function which is *analytic in (complex) frequency which is holomorphic in the lower half of the complex frequency plane*. It is worth clarifying that in Jackson's book [14] the domain of analyticity is the *upper* half-plane. This difference is just a consequence of differing conventions on how the complex function is put together. It amounts to an interchange of ϵ with ϵ^* .

In the very next para (after his Eq. (22)), Kramers shows the clear connection between this analyticity and *causality*. Thus, Kramers's 1927 paper is the clearest elucidation of the relevance of causality and analyticity to the behaviour of scattering amplitudes. Jackson, Kramers and Kronig all had to use one additional relation, Eq. (20) of Kramers, Eq. (7.113) of Jackson's book and similar relation in Kronig's later work (we shall come to it shortly), which is not a consequence of analyticity. This too will turn out to be another major ingredient in the modern theory of S-matrix. We will explain the significance of this additional relation soon (at that point, we will correct a crucial error in Jackson's version.)

Though Kronig's 1926 J.O.S.A paper had only given the dispersion relations (for X-rays) without connecting the same to either analyticity or causality, in his 1942 "Letter to the editor", mentioned above, he cited, apart from Kramers's 1927 work, another work of his own, also done in 1942. This paper, titled *The general theory of dielectric and magnetic susceptibilities* (the original is in Dutch) [21], was also published in a Dutch journal which is rather hard for outsiders to access. In this work, Kronig also elucidates the connections between dispersion relations, causality and analyticity.

Let us see these connections in some greater detail. We will first show the connection between analyticity and the dispersion relations, and then, that between analyticity and causality. Consider a complex, but analytic function $\epsilon(\Omega)$ of *complex* frequency Ω . The physical frequency ω is to be taken as the real part of Ω . The domain of analyticity is taken to be the *upper half plane* in the complex Ω -plane. It is then a straightforward application of the Cauchy residue theorem that

¹ The author thanks the Leiden University Library for making this available.

$$\epsilon(\Omega) = \frac{1}{2\pi i} \oint_{C_+} \frac{\epsilon(\Omega')}{\Omega' - \Omega} d\Omega' \quad (4.5)$$

The integral is over any closed contour C_+ that lies entirely in the upper half-plane and the integration is in the counter-clockwise direction. Let us take this contour to run along the $\omega(\text{Re } \Omega)$ -axis from $-\infty + i\delta$ to $\infty + i\delta$ and be closed by a semicircle at very large $|\Omega|$. δ is an arbitrarily small positive and real quantity. At this point, some crucial assumptions are made. One of the most important ones is the behaviour of $\epsilon(\Omega)$ as $|\Omega|$ becomes very large. It is assumed that it falls off sufficiently fast to make the contributions to the contour integral from the large semi-circle vanish. It is very important to understand the nuances of such boundary conditions. In physics, we do encounter quantities that have large fall-offs as some physical quantities like frequency, energy, etc., take on very large values. But here, the fall-off is with respect to the magnitude of a *complexified frequency*. The latter has no clear observable meaning of its own. It should be appreciated that ϵ is not only complex because of making a pair of real functions into one complex function but also because each of these has been extended into the complex plane of frequencies. That amounts to a continuation. Often, such continuations have no uniqueness to them. The boundary conditions place restrictions on such continuations, and as we shall see as we go along, sometimes they make the continuations unique. The existence and uniqueness of continuations is technically a very subtle affair, and a deep knowledge of complex analysis is necessary.

If the boundary conditions are fulfilled, the contribution from the semicircle vanishes and one is left with

$$\epsilon(\Omega) = \frac{1}{2\pi i} \int_{-\infty}^{+\infty} \frac{\epsilon(\omega')}{\omega' - \Omega} d\omega' \quad (4.6)$$

where use has been made of $\Omega' = \omega' + i\delta$, with real ω' , and the integration variable changed to ω' . Evaluating the above at $\Omega = \omega + i\delta$ with real ω ,

$$\epsilon(\omega) = \frac{1}{2\pi i} \int_{-\infty}^{+\infty} \frac{\epsilon(\omega')}{\omega' - \omega - i\delta} d\omega' \quad (4.7)$$

On making use of the well-known identity

$$\frac{1}{\omega' - \omega - i\delta} = P \frac{1}{\omega' - \omega} + i\pi \delta(\omega' - \omega) \quad (4.8)$$

and making some simple rearrangements, one gets the final form of the dispersion relations in this case

$$\begin{aligned} \text{Re } \epsilon(\omega) &= \frac{1}{\pi} P \int_{-\infty}^{+\infty} \frac{\text{Im } \epsilon(\omega')}{\omega' - \omega} \\ \text{Im } \epsilon(\omega) &= -\frac{1}{\pi} P \int_{-\infty}^{+\infty} \frac{\text{Re } \epsilon(\omega')}{\omega' - \omega} \end{aligned} \quad (4.9)$$

It is important to notice that the prefactors in the above equations involve π , and not 2π . On comparing these to the Eq. (4.4) (and equivalent relations in the works of Kronig and Kramers) introduced earlier in this chapter, one sees several important differences. In those equations, the frequency ranges of integration were only over the physically meaningful *positive* frequencies. But the range of frequencies in Eq. (4.9) is over both positive and negative frequencies. Also, the equations for $\text{Im } \epsilon(\omega)$ have different ω -dependences.

This leads to the physical meaning one has to attach to ϵ for *negative* frequencies. It was Kramers who, in [15], had shown that negative frequency contributions arise in quantum theory. It turns out that in many theories, certainly in both classical and quantum theories of dispersion, there are additional relations of the type

$$\epsilon^*(\Omega) = \epsilon(-\Omega) \quad (4.10)$$

even for *complex* Ω . The relation in Jackson's book, Eq. (7.113), $\epsilon(-\Omega) = \epsilon^*(\Omega^*)$ is incorrect, as can easily be verified from his previous Eq. (7.112). Since Jackson only uses this for real $\Omega = \omega$, the error is harmless. There are several important properties of Eq. (4.10) that are worth enumerating:

- This does not follow from analyticity as can easily be checked directly from Eq. (4.6). Therefore, requiring this is over and above the requirements of analyticity.
- However, this equation is *compatible* with the requirements of analyticity, as can also be checked directly.
- This complex equation is equivalent to the pair of real Eq. (20) in [19].
- In the context of optics, these pair of real equations implies that the refractive index is an *even* function of ω , while absorption coefficient is an *odd* function of ω .

This relation was referred to as the *symmetry* condition by Kronig, and is what comes closest to the modern day *Crossing Symmetry* of the S-matrix. Using this very important relation, it is an easy exercise to show that Eq. (4.9) goes over to Eq. (4.4).

4.3 Connections to Causality

Now, we turn to the connection between analyticity (in the upper half-plane in complex frequency in the optics case) and causality. We shall look at the analysis of Kronig in [21] first, because it is simple and direct. We shall then look at the very clear exposition of this in classical electrodynamics as given in Sect. 7.10 of Jackson's book.

Kronig starts with a $X(t)$ that can be considered as a *cause*, whose *effect* is $x(t)$. Causes and effects can be vectorial, tensorial etc but it suffices to consider a one-component cause and effect to bring home the connection we are after. Following Kronig, let us consider an *impulsive* cause at time $t = 0$:

$$X(t) = \delta(t) = \frac{1}{\pi} \int_0^\infty \cos(\omega t) d\omega \quad (4.11)$$

where $\delta(t)$ is the Dirac delta function and ω is the frequency, Fourier-conjugate of time. It should be noted that only *positive* frequencies enter at this point. The effect is also Fourier-analysed:

$$x(t) = \frac{1}{\pi} \int_0^\infty (\chi_1(\omega) \cos(\omega t) + \chi_2(\omega) \sin(\omega t)) d\omega \quad (4.12)$$

Now, the physical principle of *causality* is that there can be no effect before the cause. More mathematically, if $\tau > 0$, then $x(t) = 0$ for $t = -\tau$. That is,

$$\int_0^\infty \chi_1(\omega) \cos(\omega \tau) d\omega - \int_0^\infty \chi_2(\omega) \sin(\omega \tau) d\omega = 0 \quad (4.13)$$

It immediately follows that, for $t \geq 0$,

$$\frac{1}{\pi} \int_0^\infty \chi_1(\omega) \cos(\omega t) d\omega = \frac{1}{\pi} \int_0^\infty \chi_2(\omega) \sin(\omega t) d\omega = \frac{n(t)}{2} \quad (4.14)$$

with the function $n(t)$ defined as above. Kronig called it the *response function*. $n(t)$ is obviously a *real* function. Now, let us consider the complex combination

$$\chi(\omega) = \chi_1(\omega) + i \chi_2(\omega) \quad (4.15)$$

Then, Eq. (4.14) is indeed a relation between the real and imaginary parts $\chi_1(\omega)$, $\chi_2(\omega)$ of $\chi(\omega)$. That is, it is a dispersion relation. It is somewhat implicit, and we shall make it more explicit shortly. Eq. (4.14) can be easily inverted to give

$$\chi_1(\omega) = \int_0^\infty n(t) \cos(\omega t) dt \quad \chi_2(\omega) = \int_0^\infty n(t) \sin(\omega t) dt \quad (4.16)$$

Hence,

$$\chi(\omega) = \int_0^\infty n(t) e^{i\omega t} dt \quad (4.17)$$

To see the emergence of dispersion relations as described, substitute, for example, the expression for $n(t)$ in terms of χ_2 in the expression for χ_1 in terms of $n(t)$:

$$\begin{aligned} \chi_1(\omega) &= \int_0^\infty dt \cos(\omega t) n(t) \\ &= \int_0^\infty dt \cos(\omega t) \frac{2}{\pi} \int_0^\infty \sin(\omega' t) \chi_2(\omega') d\omega' \\ &= \frac{2}{\pi} \int_0^\infty d\omega' \frac{\omega' \chi_2(\omega')}{\omega'^2 - \omega^2} \end{aligned} \quad (4.18)$$

and likewise

$$\chi_2(\omega) = \frac{2\omega}{\pi} \int_0^\infty d\omega' \frac{\chi_1(\omega')}{\omega'^2 - \omega^2} \quad (4.19)$$

These are exactly the same dispersion relations that we had before!

Because $n(t)$ is a real function, Eq. (4.17) tells us that $\chi(\Omega)$ is analytic in the upper half-plane of the complex Ω , as long as $n(t)$ is finite for all t . It also follows that $\chi(\Omega)$ satisfies the symmetry condition of Eq. (4.10).

Now we show how all this works out in classical electrodynamics. It is well known that in a *Dielectric*, the relation between the electric field $\mathbf{E}$ and the *Displacement* $\mathbf{D}$ is given by (assuming the medium is isotropic)

$$\mathbf{D}(\mathbf{x}, t) = \epsilon \mathbf{E}(\mathbf{x}, t) \quad (4.20)$$

where ϵ is known as the *Dielectric Constant*. But in a *Dispersive* medium, ϵ is frequency-dependent in the sense

$$\mathbf{D}(\mathbf{x}, \omega) = \epsilon(\omega) \mathbf{E}(\mathbf{x}, \omega) \quad (4.21)$$

This introduces a temporal nonlocality between the displacement and the electric field, which can be worked out by using Faltung's theorem. The result is easily worked out (see Eqs. (7.105), (7.106)) of Jackson's book [14]. We shall simply state it

$$\mathbf{D}(\mathbf{x}, t) = \int_{-\infty}^{\infty} d\tau G(\tau) \mathbf{E}(\mathbf{x}, t - \tau) \quad (4.22)$$

with

$$G(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \epsilon(\omega) d\omega \quad (4.23)$$

G is real. In a simple model, Jackson shows that $G(\tau)$ is proportional to $\theta(\tau)$ (see Eq. (7.110) of the book). In fact, this is generically true and is nothing but a statement of causality. In the light of this, Eq. (4.22) takes the form

$$\mathbf{D}(\mathbf{x}, t) = \int_0^\infty d\tau G(\tau) \mathbf{E}(\mathbf{x}, t - \tau) \quad (4.24)$$

Equation (4.23) can be inverted to yield

$$\epsilon(\omega) = \int_0^\infty G(\tau) e^{i\omega\tau} d\tau \quad (4.25)$$

As of now, Eq. (4.25) is for real frequencies only. But its analytic continuation to complex Ω is straightforward

$$\epsilon(\Omega) = \int_0^\infty G(\tau) e^{i\Omega\tau} d\tau \quad (4.26)$$

From this representation, it immediately follows that

- $\epsilon(\Omega)$ is analytic in the upper-half Ω -plane
- $\epsilon^*(\Omega) = \epsilon(-\Omega)$

The dispersion relations follow as a consequence of these.

With this, the requirements for the S-matrix have been enlarged from Heisenberg's unitarity, Lorentz invariance, and restricted form of analyticity to unitarity, Lorentz-invariance, causality and the consequent larger analyticity, i.e. dispersion relations and crossing symmetry. Next, we discuss how these notions are to be carried over to generic relativistic quantum field theories.

4.4 Causality and Analyticity in Non-relativistic Quantum Mechanics

Before considering RQFT, we briefly review the situation in nonrelativistic quantum mechanics. QM raises many foundational questions about causality, and notions like *arrival time*, etc., have to be dealt with extreme care. See, for example, [22] for various aspects of time in QM. The outcomes of measurements being random, causal descriptions on an event-to-event basis are not there. One can take a heuristic attitude to such difficulties. Schutzer and Tiomno [23] investigated the connection between causality and analyticity of S-matrix. They formulated causality in much the same way as Kronig did [21], i.e. no output before input. Specifically, they assumed no scattered wavepacket before the incident wavepacket had hit the scatterer. They showed that the poles of the scattering matrix all lay in the lower half of the complex energy plane.

N.G. van Kampen did a more careful analysis [24]. He pointed to a fundamental difficulty in the formulation of the causality principle. His incisive observation was that in non-relativistic QM, energies being positive, no wave packet could be constructed that would vanish for all $t < 0$. He could circumvent these problems for the case of scattering of electromagnetic waves. In this first paper, he considered the scattering of electromagnetic fields by finite, spherically symmetric scatterers. Once again, finiteness of the scatterers was essential for an unambiguous statement of the causality principle. It is very instructive to read Sect. III *The Causality Condition* to appreciate the nuances of the problem. van Kampen was also able to show that the S-matrix is analytic with singularities only in the lower half-plane in complex energy.

In his second paper [25], van Kampen looks at the scattering for non-relativistic particles. Contrary to naive expectations, he points out that proper formulations of causality are actually more difficult in this case. This has to do with the fact that for electromagnetic scattering, propagation was limited by light velocity and therefore by analysing at distances sufficiently far off from the scatterer, the aforementioned difficulties could be handled. But for scattering of non-relativistic particles, there is no such limitation to signal velocities. Therefore, a new and sufficiently complicated

criterion for causality had to be invoked; this had to be formulated in terms of probabilities. The criterion adopted by van Kampen was *Probability of finding an outgoing particle at distance r_1 prior to t_1 can not be greater than the probability of finding an incoming particle under the same conditions*. This too becomes a difficult criterion to adopt as probabilities can not be split into incoming and outgoing probabilities due to interference. We refer the reader to the paper for more details. van Kampen finds that causality alone is not sufficient to lead to dispersion relations. However, imposing the symmetry conditions $S(-p) = S^*(p)$ leads to Kramers-Kronig-type relations.

John Toll in his thesis of 1952, but published in 1956, describes in detail various foundational issues wrt dispersion relations. He too discusses the case of electromagnetic scattering. Though his approach is very similar to that of van Kampen, many things are worked out in explicit detail. Therefore, this thesis should be valuable to all those wishing to gain both a proper perspective as well as expertise in the technicalities [26].

4.5 Microcausality

The crucial link between causality and analyticity in Kramers-Kronig approaches was the mathematical step of Fourier transformation. This will also be the starting point of establishing a similar link between analyticity and causality in relativistic quantum field theories too. Heuristically speaking, as we move from quantum mechanics to quantum field theory, time t is replaced by the four-dimensional x^μ as the label for the degrees of freedom. This is already evident in classical electrodynamics too, but for the purposes of our interest, it sufficed to consider only the Fourier transformation between time and frequency.

So, the natural first question is that of formulating the causality principle in RQFT, and then obtain scattering amplitudes as appropriate Fourier transforms. Before delving deeper into this, it is important to raise some issues about causality. The question is whether every element of a theory has to be in accordance with causality. In relativistic theories, one such expectation would be that there should not be any *acausal* or faster-than-light propagation. Another aspect would be that signals should only propagate to the future and not to the past. When every aspect of a theory respects such criteria, one could call it a *strictly causal* theory.

What complicates matters is that such strict causality may not be necessary for the observable aspects of the entire theory to be causal. A famous example is the Feynman-Wheeler absorber theory [27], where both advanced and retarded propagations occur. Another example is Dirac's classical theory of the electron [28]. In classical electrodynamics itself, in the *Coulomb Gauge*, the scalar potential propagates instantly. Nevertheless, the whole theory is causal, as it should be. Exactly how this comes about, is fascinating to figure out! The reader is recommended to see p. 223 of [14].

Though it may appear that the passage from QM to RQFT is a technical one involving a passage from *finite* number of degrees of freedom to infinitely many

degrees of freedom, this passage introduces many conceptual difficulties too. It is not our aim to go into such details, notwithstanding their importance. We will simply mention some salient aspects and refer the interested reader to a few very good books on Quantum Field Theories. The first point to grasp is that in RQFT, $\phi(x)$ is not a function of x , but is a quantum mechanical degree of freedom labelled by x . It is operator-valued. Furthermore, $\phi(\mathbf{x}, t)$ and $\phi(\mathbf{x}', t)$ are *independent* even if $\mathbf{x}$ is infinitesimally close to $\mathbf{x}'$. Among other things, this means the *derivatives* of quantum fields have to be handled with great care. The quantum fields are very singular objects, mathematically speaking. This often necessitates the use of so-called *smeared fields* (see Barton's book [29]):

$$\phi_f(t) = \int d\mathbf{x} f(\mathbf{x}) \phi(\mathbf{x}, t) \quad (4.27)$$

Such issues already arise in QM itself. For example, plane wave solutions are not normalizable and use of wave packets is necessary for careful treatments. Even more crucial differences with QM arise regarding the existence of various important *Unitary Transformations*. Just as in QM, in QFT also one has various “pictures”, Schrödinger picture where states are *time-dependent* but operators are time-independent, Heisenberg picture where operators are time-dependent but states are not, and, the interaction picture where both states and operators are time-dependent but with time dependence dictated solely by the *interaction Hamiltonian*. In QM, there are well-defined unitary transformations that relate one representation to another. In QFTs, one of the landmark results of Rudolf Haag called *Haag's Theorem* (see the last chapter of [29]) denies their existence. As unitary transformations of the type $U(-\infty, t)$ are crucial for the construction of S-matrix, Haag's theorem appears to be a roadblock. Most practitioners assume the existence of such unitary transformations and carry on; at least in perturbative approaches like in QED, such pragmatic approaches have nevertheless been very successful. Yet another feature of QFT not present in QM is that operator relations make sense only as matrix elements between physical states.

In addition to the book by Barton mentioned above, the following books are highly recommended: Steven Weinberg [30], Bjorken and Drell [31], Gasiorowicz [32], Itzykson and Zuber [33], Barton's monograph on Dispersion Relations [34], Rudolf Haag's book on “Local Physics” [35] and Bogoliubov and Shirkov's Introduction to theory of Quantized Fields [36].

The most powerful, and “user friendly”, formulation of causality in RQFT is the so-called *Microcausality Principle*. To illustrate this, let us start with the simplest of possible fields, namely, a scalar field. Microcausality in that case is stated mathematically as

$$[\phi(x), \phi(y)] = 0 \quad (x - y)^2 > 0 \quad (4.28)$$

Here x, y are four-vectors, and the choice of metric is $(+, +, +, -)$, so that the above condition states that the operator fields (in the Heisenberg representation) *commute* for *space-like* separations. The physical interpretation given is that measurements of the $\phi(x)$ should not in anyway disturb the measurement of $\phi(y)$ [30] (p. 198). In the case of the electromagnetic field, the observability of the physical *field strengths*

was analysed at great depth by Bohr and Rosenfeld [37] which clarified the meaning of commutation relations such as Eq.(4.28) in the light of physical measurement processes.

At first sight, these relations look the same as the *Canonical Equal Time Commutation Relations(ETCR)*:

$$[\phi(\mathbf{x}, t), \phi(\mathbf{y}, t)] = 0 \quad (4.29)$$

And a doubt may arise as to why give a separate status to the microcausality relations! They are, of course, not the same. While ETCR follows from microcausality, the other way is not true. The relevant differences are important to understand. ETCR counts all the independent degrees of freedom at any instant, the same as in QM too. In microcausality, commutativity includes d.o.f. at different times too. If ETCR are formulated at time t , fields at t' such that they are space-like separated from the field at the origin are not independent.

As Weinberg points out, similar microcausality relations are postulated for *Fermionic Fields* too (p. 198 [30]):

$$\{\psi(x), \psi(y)\} = 0 \quad (x - y)^2 > 0 \quad (4.30)$$

where $\{A, B\} = AB + BA$ is the *anticommutator* of A, B . Unlike the electromagnetic field strengths (not, however, the vector potentials), the Dirac fields are not measurable. What is more, the anticommutativity of space-like separated fermion fields means that their commutators do not vanish in general. A naive application of their measurability would have actually demanded that such commutators vanish. When QED is formulated in terms of vector potentials, some extend microcausality to the vector potential fields too. There too, one can raise questions about the measurability aspects as vector potentials are not observable (though many persist under the illusion that they are, thanks to wrong interpretations of the Bohm-Aharonov effect!).

The measurement aspects should equally well be applicable to all observables of the theory (made entirely out of the fields, not involving the field-momenta), and not just the canonical fields. Therefore, microcausality should more generally imply

$$[\mathcal{O}(x), \mathcal{O}(y)] = 0 \quad (x - y)^2 > 0 \quad (4.31)$$

Nevertheless, Weinberg (p. 145 [30]) points to an unusual feature of the Hamiltonian density $\mathcal{H}(x)$. He says that in its case, the commutator should vanish even for *light-like* separations, i.e.

$$[\mathcal{H}(x), \mathcal{H}(y)] = 0 \quad (x - y)^2 \geq 0 \quad (4.32)$$

Obviously, at light-like separations, measurements should not impose any restrictions. The reader is directed to Chap. 6 of Weinberg's book. A set of observables of particular importance to us are the *Currents* and microcausality for them should read

$$[j_\mu(x), j_\nu(y)] = 0 \quad (x - y)^2 > 0 \quad (4.33)$$

One way to justify the anti-commutation relations Eq.(4.30) is that they lead to commutation relations of the type Eq.(4.33), as can be easily verified (see also [32] Eq.(2.109)). But commutation relations instead of anticommutation relations for fermion fields would also have worked. A totally different justification for the anti-commutation relations for fermionic fields, and commutation relations for bosonic fields, is that only then consistent quantizations can be carried out [31]. This is the famous *Spin-Statistics Connection*. But as argued at length by Sudarshan and Duck in their book [38], no fully satisfactory proof of the Spin-Statistics connection exists. Weinberg also brings out the necessity of *antiparticles* for the consistency of microcausality (p. 199).

Whether microcausality is an essential ingredient for RQFT is a rather subtle issue. According to Barton [29], (i) it is not at all clear whether the condition is necessary (it is clear that it is sufficient) for macroscopic causality and (ii) it is still an open question whether, and how, microcausality can be relaxed without running counter to macroscopic causality and macroscopic relativistic invariance. With our present formulations of RQFT, allowing the commutators to exist outside the light cone would at once allow instantaneous signal propagations and it is not at all clear how their influence over macroscopic distances can be tamed.

Causality has been a favourite with philosophers for a very long time. The reader is referred to David Bohm's book on *Causality and Chance in Modern Physics* [39] for a good blend of philosophy and physics. For a whacky, irreverent, yet accurate look at causality, the book *Doubt and Certainty* by Tony Rothman and George Sudarshan [40] is a must-read.

References

1. J.A. Wheeler, Phys. Rev. **52**, 1107 (1937)
2. R. Oehme, *Theory of Scattering Matrix: An Introduction to Heisenberg's S-matrix*, MPI-PAE-PTH-48-82 (1982)
3. W. Heisenberg, Z. f. P **120**, 513 (1943)
4. W. Heisenberg, Z. f. P **120**, 673 (1943)
5. C. Moeller, K. Danske Vidensk. Selskab. Mat-Fys. Medd. **23**(1) (1945)
6. C. Moeller, K. Danske Vidensk. Selskab. Mat-Fys. Medd. **22**(19) (1946)
7. W. Heisenberg, Z. f. P **123**, 93 (1946)
8. S.T. Ma, Phys. Rev. **69**, 668 (1946); Phys. Rev. **71**, 195 (1946)
9. W. Heisenberg, Z. Naturforsch. **1**, 608 (1946)
10. W. Heisenberg, *Two Lectures* (Cambridge University Press, 1949)
11. W. Heisenberg, Rev. Mod. Phys. **29**, 269 (1957)
12. R.J. Eden, Proc. Roy. Soc. Lond. **A199**, 256 (1949)
13. R.J. Kronig, Physica **12**, 543 (1946)
14. J.D. Jackson, *Classical Electrodynamics*, 2nd edn. (Wiley Eastern Ltd., 1978)
15. H.A. Kramers, W. Heisenberg, Z. f. P **31**, 681 (1925)
16. R.J. Kronig, J. Opt. Soc. Am. **12**, 547 (1926)
17. R. Ladenberg, Z. f. P **4**, 451 (1921)
18. R. Ladenberg, F. Reiche, Naturwiss. **11**, 584 (1923)
19. H.A. Kramers, Atti. Congr. Intern. Fisici. Como **2**, 545 (1927)
20. H. Kallmann, H. Mark, Ann. der. Physik **82**, 585 (1927)
21. R. Kronig, Ned. T. Natuurk. **9**, 402(1942)

22. G. Muga, A. Ruschhaupt, A. del Campo (eds.), *Time in Quantum Mechanics* (Springer Publications)
23. W. Schutzer, J. Tiomno, Phys. Rev. **83**, 249 (1951)
24. N.G. van Kampen, Phys. Rev. **89**, 1072 (1953)
25. N.G. van Kampen, Phys. Rev. **91**, 1267 (1953)
26. J.S. Toll, Phys. Rev. **104**, 1760 (1956)
27. J.A. Wheeler, R.P. Feynman, Rev. Mod. Phys. **17**, 157 (1945)
28. P.A.M. Dirac, Proc. Roy. Soc. Lond. A **167**, 148 (1938)
29. G. Barton, *Introduction to Advanced Field Theory*, Interscience Tracts on Physics and Astronomy, vol. 22 (1963)
30. S. Weinberg, *The Quantum Theory of Fields -I* (Cambridge University Press)
31. J.D. Bjorken, S. Drell, *Relativistic Quantum Fields* (McGraw Hill Publishers)
32. S. Gasiorowicz, *Elementary Particle Physics* (Wiley)
33. C. Itzykson, J-B. Zuber, *Quantum Field Theory* (McGraw Hill Publishers)
34. G. Barton, *Introduction to Dispersion Techniques in Field Theory* (W.A. Benjamin Inc)
35. R. Haag, *Local Quantum Physics*, 2nd edn. (Springer Publications, 1991)
36. N.N. Bogoliubov, D.V. Shirkov, *Introduction to the Theory of Quantised Fields* (Interscience, N.Y., 1959)
37. N. Bohr, L. Rosenfeld, Kgl. Danske. Videnskab. Selskab. Mat.-fys. Medd. **12**, 8 (1933); Phys. Rev. **78**, 794 (1954)
38. I. Duck, E.C.G. Sudarshan, *Pauli and the Spin-Statistics Theorem* (World Scientific Publishers, 1997)
39. David Bohm, *Causality and Chance in Modern Physics* (Rutledge, London, 1975)
40. T. Rothman, G. Sudarshan, *Doubt and Certainty* (Scientia, 1998)



QED: S-Matrix, Causality and Analyticity

5

5.1 Development of QED

This section should be read in conjunction with Sect. 1.5 of Chap. 1. In that section, we described the genesis and completion of Quantum Electrodynamics, the first highly successful RQFT. That description, though accurate and detailed, was non-technical. Here, we take up a more technical discussion of QED with special reference to the issues of causality, analyticity and the S-matrix.

The history of Quantum Field Theory, and QED in particular, presents a peculiar situation in that many ground-breaking technical and conceptual developments took place already in the 1930s, almost immediately after the development of non-relativistic Quantum Mechanics in 1925, and the immediate development of relativistic quantum mechanics by Klein [1], Gordon [2] and Dirac [3]. Actually, Schrödinger was the first to write down and investigate what came to be known as the Klein-Gordon equation. He did this even before developing his wave mechanics, but rejected it as it produced a fine structure for the hydrogen atom at variance with experimental data (see Biedenharn [4] for details). Though Dirac discovered his equation 2 years later in 1928, the complete solution to the spectrum of hydrogen atoms in Dirac theory was solved by Gordon [5] and Darwin [6], both in the same 1928. A striking feature of this relativistic Coulomb spectrum for electrons was the *exact degeneracy* of levels with the same (n, j) quantum numbers for all values of l consistent with the given j . Thus, the levels $2s_{1/2}$ and $2p_{1/2}$ would be degenerate. This will be seen to play a big role in the development of QED.

Just on the basis of intuition, one can see why putting relativity and quantum mechanics must inevitably lead to quantum field theory. Heuristically, localizing a particle to finer and finer levels would increase its momentum, and consequently, its energy. In an interacting system, this energy can be redistributed in many ways, and $E = mc^2$ of relativity would permit such redistributions into creation of more particles. A quantum field is also one whose excitations cover an arbitrary number of

particles. Nevertheless, the *single-particle* Dirac or Klein-Gordon equations did not display this essential feature. In the Dirac case, progress came on recognizing the importance of the *negative energy solutions*. Their presence made electrons *unstable* as all positive energy electrons could make transitions to these negative energy states and disappear. To circumvent this, Dirac proposed that *all negative energy states are filled*. That electrons, being half-integer spin particles obeyed the Pauli exclusion principle was critical for the consistency of this picture. Now, a sufficiently energetic photon absorbed by a negative energy electron in the filled Dirac sea could make a transition to a positive energy real electron, leaving behind a *hole* of positive energy and positive charge. Dirac had originally suggested this state to be the proton [7]. Robert Oppenheimer pointed out that if they were protons hydrogen atoms would decay with an incredibly short lifetime [8]. Hermann Weyl [9] showed that the Dirac equation has a mathematical symmetry (in modern parlance, the charge conjugation invariance) that would require that the hole in the Dirac sea have the same mass as the electron, thus ruling out the proton interpretation. This state eventually came to be identified with new particles, *Positrons*, which were experimentally discovered by Anderson in 1932. Thus was born the concept of *antiparticles*. The important fallout is that the intuitively anticipated particle creation in relativistic quantum systems got a concrete mechanism. This was the beginning of the making of a Relativistic Quantum Field Theory (RQFT).

The Schrödinger-Klein-Gordon equation also admits negative energy solutions. Since the Pauli exclusion principle does not apply in this case, the Dirac prescription of filling all negative energy states does not work. The issues are more subtle in this case; Pauli and Weisskopf gave a resolution to this [10]. Here too the additional states got interpreted in terms of antiparticles.

Pauli and Heisenberg wrote two papers that were essentially the beginning of Quantum Field Theories [11]. In these, they laid down the foundations of the so-called *canonical quantisation* which have survived even to this day. They applied their ideas to both the Maxwell electromagnetic fields as well as the Dirac fermion fields. They also worked out various important aspects of field theories like Lorentz symmetries and Gauge invariances. They had even anticipated the divergence difficulties inherent in generic quantum field theories. Shortly before that, Dirac had started his early works on the beginnings of QED in the pioneering papers [12, 13]. The equivalence of the Dirac hole theory with quantum field theory was shown by Fock [14], and by Furry and Oppenheimer [15] in 1933 and 1934, respectively. Nevertheless, hole theory continued to be used because of its apparent intuitive appeal.

As mentioned earlier, many conceptual and technical issues that had a crucial bearing on the development of modern QED had already been initiated and well studied in the 1930s itself. The problem of electron self-energy was one of them. In one of the earliest investigations of this problem, Robert Oppenheimer in 1930 studied the electromagnetic self-energy of bound electrons [16]. In this pioneering work, Oppenheimer noted that, while self-energies of individual levels diverged, in the differences between energy levels much of the divergences cancelled. Unfortunately, the significance of this very perceptive observation was not appreciated till much later. Oppenheimer had not taken into account the influence of the negative energy

states (i.e. of positrons) on the self-energy problem. Weisskopf remedied this situation by including intermediate states containing positrons (see Ref. [63] of the first chapter of Weinberg's book [17]). In this work, Weisskopf made the remarkable observation that the inclusion of positrons led to cancellations of most divergences in the self-energy of free electrons. While the first work had only been done to leading order, Weisskopf later gave an *all order* demonstration of this in [18]. As we shall see shortly, this too proved to be crucial in the epoch-making developments of QED.

Another effect which occupied considerable attention later was the so called *Vacuum Polarization*. Here too divergences were encountered. Dirac in 1933 had suggested the idea of *Charge renormalization* to handle these divergences. The idea being very similar in spirit to that of Kramer's mass renormalization, i.e. the physically observed charges must already subsume this effect where by even the vacuum gets polarized. Already in 1935, Uehling [19] had shown that even after renormalizing the charge, a residual and observable shift of atomic levels would persist. The same year, an even more complex manifestation of the polarization of vacuum was calculated by Hans Euler and B. Kockel on the one hand [20], and, Heisenberg and Euler on the other [21]. This was the phenomenon of *light by light scattering*. Euler was Heisenberg's doctoral student and this was the topic of his dissertation. He died just a few years later in 1941, at the age of 32, as a Luftwaffe pilot. Initially, this calculation too had been thought to suffer from the problem of divergences. A year later, Kemmer and Weisskopf [22] showed that these divergences were spurious and that gauge invariance renders them finite.

As emphasized in Sect. (1.5), the development of QED was driven almost entirely by the challenges thrown up by the experiments of Lamb and Retherford [23] on the one hand, and of Foley and Kusch [24] on the other. The latter was the observation of a tiny (1 in 1000) departure from Dirac's prediction of the magnetic moment of the electron ($g_S = 2$), and the former, a tiny relative displacement of about 1030 MHz between the $2S_{1/2}$ and $2P_{1/2}$ levels of hydrogen, subsequently called the *Lamb Shift*. According to Dirac theory (more precisely, the spectrum of the hydrogen atom as given by the single-particle Dirac equation), these levels were exactly *degenerate*. That experiments may be in disagreement with the Dirac theory had been noted by Pasternak as early as 1938 [25]. For a very thorough historical and technical account of the development of QED, the reader is highly recommended to read Silvan Schweber's book on QED [26]. This also contains a wealth of highly relevant references (there are, however, some grave errors, for example, about the works by K.M. Case and M. Slotnick.). The book also gives many useful technical derivations in detail. Some of the pathbreaking original papers can also be found in the reprint collection *Quantum Electrodynamics* edited by Schwinger [27].

The revolutionary developments leading to the successful formulation of QED owe themselves to a very large number of people, as indeed is natural to science. Nevertheless, it is in order to identify a small set of people whose contributions were critical and central. That includes Hans Bethe, H.A. Kramers, Julian Schwinger, Richard Feynman, Freeman Dyson, Victor Weisskopf and S. Tomonaga (and his school), over and above the earlier pioneers that we have already discussed. Their approaches were dramatically different, both technically and conceptually. Broadly

speaking they can be separated into the (i) Schwinger-Tomonaga approach, rooted in methods of Quantum Field Theory and (ii) the Feynman approach, which started off with very intuitive particle-based ideas, but converged remarkably with the more formal-looking but physical Schwinger-Tomonaga approach. The equivalence of these apparently different approaches was demonstrated by Freeman Dyson in a very important work [28]. In another very important work, Dyson gave the S-matrix formulation of QED [29]. In that work, he established many important concepts and results, and showed how they can be established to all orders in perturbation theory. Issues of concern to us in this book, namely, causality, analyticity and the S-matrix, need to be analysed separately for these distinctly different approaches to QED. We will also point out some crucial differences between the S-matrix of Dyson, and the S-matrix as envisaged by Heisenberg.

What kicked off the modern race to QED was Hans Bethe's incisive intuition that was exceptionally on target, that the Lamb Shift was a consequence of the quantum fluctuations of the interacting electron and photon fields. Bethe's reasoning was remarkable to the point; the quantum mechanical system representing the atom should no longer be considered as *isolated*, but rather as interacting with the quantized electromagnetic field. This electromagnetic field is over and above the Coulomb field of the nucleus. The absolute necessity for such a quantized electromagnetic field can be understood, as without it, the stationary states of the atom would be stable forever. It is the interaction between the atom and the quantized electromagnetic field that spoils the stationarity of the pure atomic states. In other words, the combined state with the atom in an excited state and the electromagnetic field in its vacuum state (no photons) evolves, under the interaction, to the ground state of the atom and a one-photon excited state of the electromagnetic field. It is unfortunate that this very basic physics is not properly explained in countless number of atomic physics and quantum mechanics textbooks.

Even before Lamb's experiment, the discussions at the Shelter Island, and Bethe's non-relativistic calculation, Weisskopf had, already by 1946 assigned the problem of the shift of the energy levels to his student Bruce French. Post the Shelter Island, he intensified his investigations into this problem. On the train back from the Shelter Island conference, Bethe produced his now famous non-relativistic calculation of the Lamb shift [30]. Bethe, in this paper, credits Oppenheimer, Weisskopf and Schwinger for identifying the interaction of the electron with the radiation field as the source of the energy shift. He extended the physics behind photon emissions and absorptions to second-order effects which can cause radiation-less effects. Basically, the atom first makes a *virtual transition* to an atomic state plus a photon, and then makes a second virtual transition to a single atomic state (the original). In the process, as is well known from perturbation theory, the energy of the state is *shifted*. Again, as is already well understood in quantum mechanics, such virtual transitions need not conserve energy. Hence, the frequency of the intermediate photon is *unrestricted*. Summing over the contributions of all frequencies (i.e. from 0 to ∞), the shift *diverges*. In fact, the divergences come from both very high frequencies (ultraviolet) as well as very low frequencies (infrared). As we shall see, both played major roles in the tortuous journey to a successful formulation of QED.

Bethe was already familiar with some earlier work by H.A. Kramers on ways to handle such diverging *self energies*. Kramers had analysed the problem of diverging self-energies in classical electromagnetic theory, where the self-energy of a point charge diverges [31,32]. In that context, Kramers introduced the concept of *mass renormalization*, according to which the observed, finite, mass of the charged particle has already accounted for this *diverging* contribution, and that the mass of the charged particle in the absence of this is unobservable even in principle and hence devoid of any physical significance. Stated differently, the sum of the *bare* mass m_0 (unobservable), and the diverging self-energy δm (a travesty to use the symbol δ usually used to denote small variations!), $m_0 + \delta m$ should be identified with the observed physical mass. That a diverging δm would then imply a diverging m_0 was to be of no concern, being unobservable.

Even after performing this mass renormalization, a milder, logarithmic divergence still persisted. Bethe sought to control this on the hopes that a more complete relativistic treatment would provide a natural cut-off for this divergent integral. He set the upper limit for this at $k = mc^2$. After a few more reasonable sounding approximations, he obtained the value of 1040 MHz for the Lamb shift, remarkably close to the experimental value.

Recall our earlier mention of Weisskopf's work [18] wherein he showed that including positrons in intermediate states removed many of the divergences. In the non-relativistic calculations of Bethe, the divergences were *linear* before mass renormalization which tamed them to logarithmic divergences. According to Bethe (as stated in his paper [30]), Schwinger and Weisskopf were the first to suggest that hole theory (meaning inclusion of positrons) must be used to obtain convergence. The reasoning was that with the inclusion of positrons, the linear divergences of the non-relativistic theory would be tamed to logarithmic divergences, and Kramer's mass renormalization would render them finite (strictly speaking, in the limit, the cut-off is removed). Indeed, the natural cut-off provided by the relativistic theory is infinity itself.

Soon after the Shelter Island meeting, Dyson joined Cornell in 1947 and started to work with Bethe, who asked him to investigate a fully relativistic calculation of the Lamb Shift for the hypothetical case of the electrons being spin-0. Dyson was quite well versed in field theory by then, having studied Gregor Wentzel's highly influential book on the subject [33]. He also made use of the Pauli-Weisskopf treatment of negative energy states of the Klein-Gordon equation that we mentioned a short while ago [10] as well as Weisskopf's pioneering works on self-energy calculations [34,35] to take care of intermediate states with antiparticles correctly. He gave a novel, physically transparent calculation of the vacuum polarization contributions. Following Weisskopf he had separated the self-energy into *transverse* and *longitudinal* parts. Dyson's results were published in [36] (submitted in December 1947). It is a highly readable and complete account of a first fully relativistic treatment of the Lamb Shift. Unlike the Dirac case, the divergences before mass renormalization are now *quadratic*. While Dyson found subtractions (read mass renormalization) to reduce this degree of divergence, he found the longitudinal and transverse contributions to be still logarithmically divergent by themselves, but almost cancelling

each other. In the end, he did obtain a convergent result as was expected of a fully relativistic treatment. It is somewhat confusing that Dyson still talks of cut-off even after obtaining convergent results. Though his results were convergent, it was far from obvious if they were *unambiguous*. Despite the relativistic treatment, many intermediate steps were not manifestly covariant.

Depending on the details of subtraction, he obtained 1003–1040 MHz for the Lamb-like shift for the difference between 2s and 2p levels. The exact physical significance of these numbers is not at all clear. Dyson interpreted the fact they are so close to Bethe's non-relativistic values (where spin effects should only make small corrections) to mean the essential correctness of Bethe's ideas about self-energy effects. This is far from clear as Lamb shift is a small residual effect among much larger effects. Another important point to consider is that, while the Dirac spectrum is labelled by four quantum numbers, the Klein-Gordon spectrum is labelled by only three. In fact, as first worked out by Schrödinger, the 2s and 2p levels for the relativistic spin-0 atom are not even degenerate, being separated by fine-structure-like separations (curiously, Dyson makes no mention of Schrödinger's spectrum calculations!). Nevertheless, Dyson had produced a fully relativistic calculation of the level shift in an atomic system whose source was the interaction of the atom with the quantized radiation field.

Literally, a race had begun to provide a sound theoretical basis for the shift observed by [23]. Interestingly, one part of this intellectual race happened in America, led by Weisskopf and his student Bruce French, by Lamb himself along with Norman Kroll, by Julian Schwinger, and by Richard Feynman. In Japan, under the leadership of Nishina, Tomonaga and their collaborators, a similar race had been going on, completely unaware of what was going on in America. The two nations were separated by distance and war, but united in science. By the spring of 1948, several claims were made for a fully satisfactory relativistic solution, both in Japan and in America.

Before going into these amazing developments, it is important to narrate a rather crucial interlude. After the Shelter Island meeting and Bethe's pathbreaking non-relativistic calculation of the Lamb Shift, Schwinger and Weisskopf had come to the realization that a fully relativistic treatment was what had to be done. At the Shelter Island meeting, Gregory Breit had reported on some of his theoretical work, carried out at the behest of Isidore Rabi, into what appeared to be yet another anomaly [37]. Rabi et al. [38] had observed that the hyperfine splitting of the $^2S_{1/2}$ state of hydrogen was 1421.3 MHz instead of the expected 1416.9 MHz based on the observed magnetic moment of the proton and the assumption of $g_S = 2$ for electrons. Breit had tried to work out possible explanations for this and had suggested that an additional magnetic moment to the electron, over and above the value $g_S = 2$ predicted by Dirac would be a natural explanation. Spurred by this suggestion, Rabi asked Foley and Kusch to investigate this issue more thoroughly. By combining data from gallium and sodium, Foley and Kusch showed that $g_S = 2.00244 \pm 0.00006$, a clear deviation from Dirac theory. This final result of Foley and Kusch was submitted around the end of December 1947 [24].

It is not clear what impression Breit's presentation had on Schwinger, but according to Schweber [26], around September 1947, he learnt of the Nafe, Nelson and Rabi work, from Norman Ramsey and told Ramsey that he thought he could calculate it, also remarking that it would be much harder than Bethe's nonrelativistic calculation. Around this time, Schwinger had already revisited the Bethe calculation but in a completely independent manner, making use of his hall-mark *canonical transformations* and shown that he could reproduce Bethe's results. He then turned his attention to hole-theoretic methods to calculate radiative corrections to Coulomb scattering, and also to a calculation of the Lamb Shift. Concurrently, he also decided to take a look at the magnetic moment issue. He chose his Hamiltonian approach augmented by suitable unitary transformations towards this end (for details see Chap. 7 of [26]). By the end of December 1947, just a few days apart from Foley and Kusch's submission, he announced that his theoretical calculations had yielded an additional magnetic moment of $\frac{\alpha}{2\pi}$ for the electron, in remarkable agreement with the experimental values. Foley and Kusch had to assume that $\delta g_L = 0$ and they had based it on the *Correspondence Principle* as in classical electrodynamics $g_L = 1$. This is of course a fallacious argument. They acknowledge Schwinger to the fact that his calculations had indeed yielded $\delta g_L = 0$.

The magnetic moment calculation of Schwinger had a tremendous psychological impact, pointing towards the essential correctness of the identification that the interaction of the electrons with quantized radiation field was at the crux. It also gave confidence that fully relativistic treatments of these interactions were crucial.

Weisskopf and French continued working with hole theory, and this formed French's Ph.D. thesis dissertation [39]. They were the first to obtain the correct result for the Lamb Shift. However, the publication of their result [40] took considerably longer. The reason was Weisskopf's lack of confidence as their result was not agreeing with the results of Schwinger and Feynman, both of which were in agreement at that time, but found to be wrong later. French and Weisskopf obtained the value of 1051 MHz for the shift and they did not have to make any further assumptions about the anomalous magnetic moment of the electron, which actually makes a contribution. French and Weisskopf submitted their paper as late as 10 December 1948, which was published only by 15 April 1949. Kroll and Lamb also used hole-theoretic methods and also obtained 1051 MHz for the Lamb Shift but only after assuming the Schwinger-Foley-Kusch value for the anomalous magnetic moment. Their paper [41] was submitted on 7 October 1948, and published on 1 February 1949.

Taking a purely chronological stand on the Lamb Shift calculations, we take up the work by the Tomonaga school in Japan. That would also be the right tribute to the focus and tenacity of this group against great odds of the war. That they successfully developed QED while being totally isolated from the Americans is truly inspiring. A bit of history should help in placing the achievements of the Japanese group in the correct perspective. Tomonaga is among the few (Weisskopf is the other) whose works nicely connect the works of the early 1930s with what culminated into modern QED. An excellent account of Tomonaga's works as well as his persona is to be found in Schwinger's tribute *Two shakers of physics* after Tomonaga's death [42]. As early as 1933, he, along with his mentor Nishina, worked out the conversion of gamma

rays to electron-positron pairs in the presence of a nucleus. Tomonaga was deeply influenced by Dirac's *Relativistic Quantum Mechanics* [43] (this is Dirac's early work on QED, not to be confused with his 1928 paper on the Dirac equation). In this paper, Dirac sought to demote the role of the fields in deference to those of the particles, an idea that Schwinger later criticized as a *false trail* [42]. Tomonaga showed the equivalence of Dirac's QED to the field theories of Heisenberg and Pauli. He made use of *Unitary Transformations* to this end. Unitary transformations, also called *Contact Transformations* were extensively used by both Tomonaga and Schwinger in their approaches to modern QED. These were pioneered by Bloch and Nordsieck as early as 1937 [44], and somewhat later by Pauli and Fierz [45]. After a foray into nuclear physics when he went to work with Heisenberg in 1937 Tomonaga returns to relativistic quantum field theories in general, and to QED in particular.

Right from the beginning Tomonaga advocated manifestly covariant methods. In 1943, Tomonaga published his *Relativistically Invariant Quantum Field Theory* which first appeared as a RIKEN report and in 1946 its English translation appeared in *Progress of Theoretical Physics* [46]. In this work, Tomonaga developed his now famous covariant generalization of the Schrödinger equation. The reader is encouraged to read this in the original for an exceptionally lucid discussion. The usual Schrödinger equation, which is not manifestly covariant, involves the total Hamiltonian, which is also not a covariant concept (as it transforms like the time-component of a four-vector.) In order to arrive at a manifestly covariant generalization of the Schrödinger equation, Tomonaga not only made states depend on a space-like hypersurface (in this, he was strongly influenced by many of Dirac's earlier ideas) but also used a unitary transformation to eliminate the free part of the Hamiltonian, thereby making the generalized Schrödinger equation depend only on the *invariant* interaction Hamiltonian. Schwinger too arrived at the same equation some four years later, albeit totally independently (it was Schwinger who coined the phrase *interaction representation*). It should not come as a surprise that through suitable unitary transformations, the free part of the Hamiltonian could be eliminated. In fact, the *entire* Hamiltonian can also be eliminated from the Schrödinger equation, leading to the *Heisenberg Representation*. Yang and Feldman [47] in fact formulated QED in the Heisenberg representation. In the same work, Tomonaga sought to replace the equal time commutation relations, which lacked manifest covariance, by manifestly covariant forms valid for arbitrary space-like separated events. In this form, they are identical to what we have referred to earlier as *Microcausality*. Such covariant generalizations of the equal time commutation relations also formed the basis for Schwinger's own manifestly covariant formulation of QED.

In this chapter on the S-matrix, it is curious to note how both Tomonaga and Schwinger turned to the S-matrix (Scattering-matrix) during their respective wartime works on magnetrons and ultra-short wave circuits (Tomonaga), and microwaves and waveguides (Schwinger). Both found the framework of the S-matrix as the best-suited strategy for optimum information, precisely the purpose for which Heisenberg had introduced them in the context of elementary particles! For a more incisive discussion of Tomonaga's many important contributions, see [42].

Tomonaga did get the news of the Lamb-Retherford experiment as well as Bethe's nonrelativistic calculation, although through popular media ("a popular science column of US weekly magazine" as per Tomonaga's own account). Tomonaga goes on to say that the information about Lamb Shift prompted them to begin a calculation more exact than Bethe's. The tools that Tomonaga and his collaborators used were the manifestly covariant formulation combined with the techniques of unitary transformations. They were also in control of various crucial aspects like mass renormalization, charge renormalization, invariant subtraction methods, etc. Despite the late start, by September 1948, Tomonaga and his team had successfully calculated (i) the anomalous magnetic moment of the electron, obtaining the Schwinger value of $\frac{\alpha}{2\pi}$, (ii) the Lamb shift; this was a fully covariant *quantum field theoretic* calculation which gave finite results. They quoted a value of 1076 MHz when vacuum polarization effects were totally neglected. After including the vacuum polarization contribution of -27 MHz for the S-level (the Uehling value [19]), their final value for the Lamb Shift came out to be 1049 MHz, which differed slightly from what the American groups had obtained. This, as explained in their note added in proof, was due to their having adopted Bethe's older estimates in evaluating the "log" terms. Upon using improved values, they also obtain 1051 MHz, (iii) order e^2 corrections to Coulomb scattering (Rutherford scattering). These results were presented in their paper *A Self-Consistent Subtraction method in Quantum Field Theory* [48,49]. Their manuscript was submitted on 23 September 1948, but published on 1 March 1949.

After the war, Tomonaga had sent all their papers to Oppenheimer, who suggested that Tomonaga write a summary account of all their work. This was published on a priority basis as the paper *On Infinite Field Reactions in Quantum Field Theory* in *Physical Review* [50].

We now turn to the path advocated and pioneered by Julian Schwinger. Eventually, Schwinger adopted manifestly covariant field theoretic methods which were very close both in spirit and detail to those of Tomonaga and his coworkers. Soon after the Shelter Island meeting and Bethe's non-relativistic calculation, Schwinger, as a warm-up, repeated the non-relativistic calculation, but by using his techniques of Hamiltonians and unitary transformations. After successfully reproducing Bethe's results, he decided to push further with a relativistic calculation of the Lamb Shift as also the radiative corrections to Coulomb scattering. As mentioned earlier, even while working on the Coulomb problem, he decided to apply his techniques to the problem of an electron in an *external magnetic field* and derived the anomalous magnetic moment which agreed remarkably with the experimental results. This was a clean sweep for him.

The other calculations proved trickier. About a year after the magnetic moment paper [51], he announced his results for the radiative corrections to electron scattering [52]. In a revealing footnote, Schwinger clarifies the difficulties he faced. The spin-orbit coupling to which the anomalous magnetic moment makes a contribution, was inconsistent with his earlier successful calculation, signalling a breakdown of relativistic invariance. Schwinger attributes these to the incorrect transformation properties of the electron self-energy obtained from the Hamiltonian methods. This same lack of relativistic invariance also affected his Lamb Shift calculation. He had

already mentioned this calculation in his magnetic moment paper itself [51], but without quoting a precise value for it other than to say *the values yielded by our theory differ only slightly from those conjectured by Bethe on the basis of a non-relativistic calculation, and are, in good agreement with the experiment.*

It is at this point that Schwinger decided to base his further work on manifestly covariant methods, and developed the same covariant generalization of the Schrödinger equation as the one derived by Tomonaga, and also to work with covariant generalization of the equal time commutation relations. While the covariant calculations removed the difficulties with the incorrect spin-orbit coupling, there still were many difficulties. The divergent integrals had to be manipulated with great care, and Pauli was particularly critical of Schwinger's handling of the singular integrals. In the process, Pauli and Villars proposed their *Invariant Regularization* [53]. But Schwinger kept improving his techniques faster than Pauli was piling his criticisms! Interestingly, in his paper *Quantum Electrodynamics III* [54], Schwinger makes no reference to the Pauli-Villars paper. But the Pauli-Villars paper had a major influence on manifestly invariant calculations that followed. The other two papers of Schwinger [55,56] laid the foundations for the subject.

Even with the manifestly covariant approach, there were still some hiccups. Both in Bethe's first non-relativistic calculation, as well as in subsequent relativistic calculations, both *ultraviolet* and *infrared* divergences occurred, and both had to be eliminated to obtain finite expressions that could be compared with experiment. The same problem had been encountered by Feynman in his radically different approach to QED (to be discussed next), and both of them had handled it in essentially same manner by giving the photon a mass. Of course, in the end, this mass had to be removed, maintaining relativistic invariance. It turns out (see [26] p. 244) that both Schwinger and Feynman had made the same error in "joining" their calculations to Bethe's calculation. This was the source of disagreement between French and Weisskopf on one hand and Feynman and Schwinger on the other. French was the one who resolved the difficulty. Of course, before Feynman and Schwinger got to agree with each other, and disagree with French and Weisskopf, they had disagreed for other reasons (for Schwinger the spin-orbit problem, and, for Feynman the issue of whether vacuum polarization had physical significance or not). Such was the tortuous path to the final summit of QED.

Now we turn to Feynman's approach to QED. This was dramatically different from the approaches we have discussed so far. It was neither based on hole theory nor on quantum field theory. It was the culmination of years of the very original, off-beat approach of Feynman to all of physics. It all started with the program he started with John Wheeler, his advisor, to rid physics of self-interactions, equivalently, self-energies. Yet, the very fact that accelerated charges radiate must result in radiation reaction on the charged body. First, at a classical level, Feynman and Wheeler allowed both advanced and retarded potentials, thus giving up causality in a naive sense. Of course, in its final, observable manifestations, causality is fully preserved. Among the various issues Feynman faced was the occurrence of divergences on the light cone, and ways to regulate such divergences. He made use of this experience in handling the divergences he encountered later in the QED calculations.

The Feynman-Wheeler theory could be described by an *Action* but not by a Hamiltonian. This was due to the fact that dynamical variables at more than one instant were needed in the description. Therefore, Feynman was looking for a formulation of quantum mechanics that would be based directly on the action. The lore goes that Herbert Jehle brought to his attention Dirac's paper on the Lagrangean in Quantum Mechanics [57], and Feynman developed his famous *Path Integral Formulation* from there. An impression has been created that Dirac provided only some preliminary, but important ideas. As this author has shown, based on careful scrutiny of Dirac's writings, Dirac in fact had the entire path integral fully developed; what's more, he had the prescriptions for a much more general framework than Feynman's [58]. A very clear exposition of the Feynman-Wheeler theory and the genesis of the path integral formulation can be found in Feynman's thesis [59], as well as in the paper [60]. Schweber's book [26] also discusses things in great detail. Later on, Feynman extended his ideas to quantum fields also, though he encountered many difficulties in the case of fermionic fields. The Lagrangean(Action)-based approaches are better suited for manifestly covariant treatments, whose importance should already be obvious by now. It should be added at this point that Schwinger arrived at a (quantum)Lagrangean formulation of his own based on his *Quantum Action Principle* [61].

In many ways, the crucial ingredient to Feynman's approach to QED is his paper on the *The Theory of Positrons* [62]. In this work, Feynman gives a completely radical and new interpretation of positrons as compared to the hole-theoretic description. There is no longer the need for the infinitely filled Dirac sea and a picture where by positrons are vacancies in this sea. Feynman bases his consideration on the so-called *Green functions*, also called *Kernels*. By demanding that only positive energy solutions of the Dirac equation should propagate forwards in time, he shows that one is forced to the conclusion that negative energy states **must** propagate *backwards* in time! These are interpreted as positrons. Feynman was very careful to emphasize that causality was not violated by the backward propagation in time as only the *unphysical* negative energy states were so propagated. Feynman treated problems of scattering by external potentials of electrons and positrons. In one go, he could treat problems of not only scattering but also of pair creation, pair annihilation, etc.

The broad picture was that particles (i.e. electrons and positrons) would propagate from some initial space-time point to another space-time point where the external potential would act on them, and subsequently, they would propagate to their final space-time point. The action of the potential could be to scatter, both forwards and backwards in time. This could be generalized to arbitrary number of interactions with the potential, if a perturbative treatment in the potential made sense. The propagation from one space-time point to another is treated as if the particles were free. The great advantage of Feynman's approach was that it became possible to visualize even very complex processes, notwithstanding the frequent dangers such visualizations bring with them. Stated differently, processes could be represented *graphically* with rules for computing the quantum amplitudes for them. At this stage, the diagrams so obtained should actually be called *Space-time Diagrams* where a straight line is a symbolic representative of the state being propagated from one point to another. The

passage to the popular *Feynman Diagrams* requires first a passage to the momentum representation.

Feynman, and also Dyson, saw great parallels with the S-matrix of Heisenberg. We shall make some pertinent remarks about this later on. The passage from the above-mentioned space-time diagrams to those for the calculation of S-matrix elements, popularly called *Feynman Rules*, requires going to momentum representation and the process of *amputation of external legs*. These details are standard textbook material now.

At this point, it is worth pointing out some misconceptions. None of the above really requires the path integral representation. What are crucial are the various *propagators* or *Kernels*. This, as demonstrated in Feynman's Positron paper, does not require the path integral representation at all. What is required are the single particle solutions to the relativistic equations, their properties like completeness, etc. The covariant aspects can be demonstrated explicitly. Feynman incorporated all these ideas into his treatment of QED in [63] and the validity of the rules were established in [64]. Feynman obtains the photon propagator through a relativistic, causal generalization of the Coulomb potential. It turns out to be the well-known $\delta_+(s^2)$. His approach to the divergence difficulties of QED were inspired by how such a light-cone singularity was tamed in the Feynman-Wheeler theory [65]. It should be appreciated that Feynman was adopting a predominantly *particle-picture* unlike Schwinger and Tomonaga whose formulations were grounded in relativistic quantum field theory (RQFT).

Now, we address the way Feynman tackled the QED calculations. Armed with his repertoire of rules, he undertook a careful study of the energy shift issue. By the end of January 1948, he had made rapid and considerable progress. At the New York APS meeting of January 1948, he announced that he had made all the calculations that Schwinger had made and "Agreed with Schwinger" (see [26] p. 427). At a time when Schwinger had encountered the discrepancy about the magnetic moment of the electron, Feynman had obtained the correct $\frac{\alpha}{2\pi}$ even in the energy shift calculations. Though he was satisfied with the efficacy of his positron picture, he says that "...but nobody believes me because I have not got everything complete yet.." ([26] p. 428). He was referring to the fact that he had not completed the Lamb Shift calculation yet. Though his rules showed how to calculate vacuum polarization, Feynman had felt great unease about the physical relevance of vacuum polarization, though in hindsight this is extremely surprising. To quote him, "Polarization of vacuum still remains somewhat of a puzzle" ([26] p. 428). A testimony to his scepticism is the footnote 18 to his paper [63]; Feynman expresses the view that "It would be interesting to calculate the Lamb Shift accurately enough to be sure that the 20 Megacycles expected from vacuum polarization are actually present".

Because of this, he initially did not include the vacuum polarization contribution to the Lamb Shift, leading to a disagreement with Schwinger. Subsequently, after Schwinger had fixed the anomalous magnetic moment issue by adopting manifestly covariant methods, Feynman too changed his values by taking into account the vacuum polarization contribution. With that Schwinger and Feynman had reached an agreement, but with both of them differing from French and Weisskopf. Feynman

and Schwinger had both made identical errors in the treatment of the infrared difficulties and neither of them could easily pinpoint the reasons for the discrepancy. This was finally resolved by French, and after correcting for it, all of them agreed on the Lamb Shift value.

5.2 Dyson Equivalence Proof

Given that the approaches of Schwinger-Tomonaga on the one side, and of Feynman on the other, were so dramatically different, it was natural to surmise whether they were describing the same natural phenomenon after all. In view of the equivalence proofs of Fock [14] and Furry and Oppenheimer [15], the hole-theoretic approaches can be subsumed under field theory. Dyson, in 1948, proved the complete equivalence of the approaches of Feynman, Schwinger and Tomonaga [28]. It is a very clear and readable paper. The essence of Dyson's proof is explained here.

The starting point for Dyson's proof is the Schwinger-Tomonaga covariant generalization of the Schrödinger equation (for quantum fields)

$$i\hbar c \frac{\partial \Psi(\sigma)}{\partial \sigma(x_0)} = H_I(x_0) \Psi(\sigma) \quad (5.1)$$

As already explained, the state $\Psi(\sigma)$ in the Tomonaga-Schwinger theories is a *functional* of the space-like hypersurface σ . The meaning of this equation is lucidly explained in [46,61]. x_0 is a point on σ at which the derivative is evaluated. Dyson also explains the operational meaning of this equation. Dyson considers a sequence of space-like surfaces $\sigma_0, \sigma_1, \dots$ filling the whole of space-time. The choice of such a one-parameter family of space-like hypersurfaces is called a *foliation*, and this can be done for arbitrary space-times by the so-called ADM (Arnowitt, Deser, Misner) construction. Dyson orders these space-like surfaces in such a way that σ_1 is slightly to the past of σ_0 , σ_2 slightly to the past of σ_1 , etc. Dyson further introduces

$$\int_{\sigma_1}^{\sigma_0} H_I(x) dx \quad (5.2)$$

to be the integral of the *invariant* interaction Hamiltonian density over the four-dimensional volume bounded by the hypersurfaces σ_1, σ_0 . He then constructs the operator $U(\sigma)$ (details are given in the paper) which satisfies

$$i\hbar c \frac{\partial U(\sigma)}{\partial \sigma(x_0)} = H_I(x_0) U(\sigma) \quad (5.3)$$

The general solution of Eq. (5.1) is

$$\Psi(\sigma) = U(\sigma) \Psi_0 \quad (5.4)$$

where Ψ_0 is any vector independent of σ . By its very construction, $U(\sigma)$ satisfies the initial condition (this important aspect is not explicitly stressed in Dyson's paper)

$$U(-\infty) = 1 \quad (5.5)$$

where by $-\infty$ one means the hypersurface in the infinite past. Because of this

$$\Psi_0 = \Psi(-\infty) \quad (5.6)$$

Putting everything together,

$$\Psi(\infty) = U(\infty) \Psi(-\infty) \quad (5.7)$$

In other words, $U(\infty)$ transforms the state of the system in the infinite past to the state of the system in the infinite future. By definition, it is the *scattering matrix* or S-matrix as christened by Heisenberg, as we have already discussed. As noted then, the notion of the Scattering Matrix already existed in Quantum Mechanics, now it has been elevated to quantum field theory. Within perturbation theory, Dyson obtains the form

$$S = U(\infty) = \sum_{n=0}^{\infty} \frac{(-i\hbar c)^n}{n!} \prod_{m=1}^m \left[\int_{-\infty}^{\infty} dx_m \right] P(H_I(x_1) H_I(x_2) \dots H_I(x_m)) \quad (5.8)$$

where P denotes what is currently called the *Time Ordered Product*. A minor confusion arises here as the time-ordered products have a Lorentz invariant meaning only if the points are relatively space-like, and in the above expression the points $(x_1, x_2, \dots, x_m)$ need not be so. But by its very construction $U(\sigma)$ is a covariant object. It should be noted that the above expression is exactly what one obtains by the more conventional and non-covariant Hamiltonian methods. But the advantage with Dyson's derivation from the Tomonaga-Schwinger equation is that at every step manifest covariance is maintained. This underscores the important point that non-covariant Hamiltonian methods do not violate relativistic invariance, but are not manifestly so. The reason Tomonaga and Schwinger insisted on manifest covariant methods is to avoid pitfalls of upsetting covariance by approximations and uncontrolled intermediate steps. Despite all that care, the occurrence of divergences makes the ground slippery as extracting finite answers for them must be done respecting covariance.

After obtaining this perturbative representation, Dyson goes on to show how to obtain matrix elements between arbitrary states of free electrons, positrons and photons. His treatment is pretty transparent and we leave it to the reader to look up his paper for the details. It essentially consists in representing the initial and final states as suitable creation operators acting on the vacuum state, representing the fields occurring in $H_I(x)$ through *mode expansions* as linear combinations of creation and annihilation operators, and finally use the commutation relations between creation and annihilation operators for bosonic fields (anticommutation relations

for fermionic fields) to move all creation operators to the left and all annihilation operators to the right. These steps were later systematized into what is now called *Wick's Theorem* [66].

One important intermediate step has not been emphasized sufficiently by Dyson. The mode expansions are applicable only for free fields, but the fields occurring in $H_I(x)$ are the interacting fields. So, the needed intermediate steps are to perturbatively expand the interacting fields in terms of free fields, so to any desired order in perturbation theory all terms in the S-matrix are expressed in terms of free fields. After that, the procedures described by Dyson can be applied.

Dyson is also able to reconstruct all the graphical rules of Feynman, thus explicitly demonstrating the said equivalence. In fact, as stressed by Dyson, certain signs (crucial for the right interference between terms of Feynman's diagrammatic method) arising out of the spin-statistics connection, are automatically determined in the light of the equivalence proof. In the Schwinger-Tomonaga theories, this is automatically taken care of by the commutation and anticommutation properties of the fields. Dyson points out that the entire Feynman view was particle oriented and the fermion field that codifies the electrons and positrons in one go was not realized upfront. The equivalence proof also made it clear that Schwinger's approach would become too unwieldy in comparison with the diagrammatic approach as the complexity of the processes grew.

It is clear that Dyson's methods can be used for any RQFT as long as the fields, their statistics and the interaction Hamiltonian are specified. And in effect, his paper gives a method for working out the Feynman rules for any field theory. But determining the field content and their statistics (yes, even that), and the interaction Hamiltonian is not always straightforward. In non-abelian gauge theories, the field content is different in different gauges, and the statistics of scalar fields could be fermionic (ghosts), etc.

5.3 S-Matrix in QED

We finally come to the S-matrix in QED. All the major efforts like understanding Lamb Shift, magnetic moment of the electron or radiative corrections to Coulomb scattering, etc., were focused towards a few observable aspects. The S-matrix as envisaged by Heisenberg coded the entire observable content of a theory. While proving the equivalence between the Tomonaga-Schwinger and Feynman approaches, Dyson sought to understand the equivalence in its broadest sense, and not just in the context of these few observables. Having formally (or in perturbation theory) the S-matrix operator, he next turned his attention to a systematic determination of the S-matrix for arbitrary states of electrons, positrons and photons [29].

This was to be to arbitrary (at least in principle) orders in perturbation theory. The paper is quite a tour de force. He showed how to determine such things as the superficial degrees of divergence by power counting, regularizing the divergences, renormalization, etc., to all orders. He showed that all divergences can be subsumed into observable charge, observable mass. That the only diagrams that are divergent are the self-energy for both electrons and photons (vacuum polarization) and the

vertex function. His method depended on curing the divergences iteratively, i.e. in any divergent diagram, one first fixes the divergences of all subgraphs. Strictly speaking, he did not solve the problem of the so-called *Overlapping Divergences*. Without that, no claim could have been made about a physically observable S-matrix. But he gave enough indications on how it may be handled, and 2 years later Salam solved the problem of overlapping divergences [67].

Dyson did not address the issues of whether such a perturbative series for the S-matrix converges, say, even for not very complicated processes. In fact, with each order, the energy dependence grows (problem of large logarithms). Even if the energies are moderate enough that this does not become a problem, the series is *asymptotic*. On physics grounds too, having an all order result may not be that important. This is because the underlying theory itself may change by the time a particular level of accuracy is reached. Indeed, once energy-scale approaches or exceeds the electroweak scale of 250 GeV, QED has to be replaced by the standard model. Vacuum polarization starts getting important contributions from the theory of quarks and gluons. Schwinger was rather emphatic about it “I am not sure I was at all interested in the mathematical question of convergence to all order. I don’t think that is a physical question” [26].

5.4 Causality, Analyticity, and S-Matrix in Non-perturbative RQFT

In this section, we wish to place the QED S-matrix and causality in QED in the broader perspective put in place by Heisenberg. In the manifestly covariant formulations of Tomonaga and Schwinger, the conventional equal time commutation relations of the Hamiltonian formulations

$$[\phi(\mathbf{x}, t), \phi(\mathbf{x}', t)] = 0 \quad (5.9)$$

are replaced by the manifestly covariant generalizations

$$[\phi(x), \phi(x')] = 0 \quad (5.10)$$

For the sake of simplicity, only the scalar case has been shown, but generalizations to other important fields like the Dirac field and the electromagnetic field were also discussed by Tomonaga and Schwinger. But, as already pointed out, the measurement aspects are not immediately evident in their cases (for the electromagnetic field, when formulated in terms of vector potentials). But subsequently, these microcausality relations are only made use for quantization of the fields. They played no obvious role in determining the analytic properties of the S-matrix. In fact, in neither the works on Lamb Shift and other observables nor in the expositions of QED in general, the analyticity perspective hardly seems to have played any role. In Dyson’s construction of the QED S-matrix also analyticity considerations hardly played any role. None of the above explored the direct relation between analyticity and (micro)causality. Since everything was based on perturbation theory, the causal aspects of the theory

were accounted for by the causal nature of the free propagators. Even in the radical Feynman Positron theory, where backward propagations in time played a crucial role, all the electron and positron propagators were causal. Therefore, one sees that the Heisenberg program of trying to construct the S-matrix from general principles of Lorentz Invariance, Unitarity and Causality (after the proposal by Kronig, as already discussed), was neither implemented nor even illustrated in QED.

The Heisenberg S-matrix envisaged bound states also to appear as poles. The Dyson S-matrix is certainly not able to realize this, being perturbative in nature. How to extend his schema to accommodate bound states is certainly not a straightforward manner, though it appears that subsequently Dyson did develop techniques to tackle the bound state issue (see [26], chapter on Dyson). In this respect, quantum mechanics appears to fare much better. As shown by Singh [68], in the Regge-pole picture (more on this later) of Coulomb scattering, the hydrogen atom spectrum indeed occurs as poles.

QED was used to check various analyticity properties subsequently (see Eden et al. book *The Analytic S-matrix* [69]). Actually, there seems to be another deep issue with regard to the QED S-matrix. S-matrix elements in field theory have to be evaluated between the so-called *Asymptotic States* (technical details will be provided in the chapter on non-perturbative QFT) and when there are massless particles exchanged, as in QED or Gravitation, asymptotically free states do not exist. In fact Eden et al. [69] explicitly mention this difficulty. Nevertheless, there are many discussions of the S-matrix even in such cases.

In the coming chapters, we take a careful look at all these issues non-perturbatively.

References

1. O. Klein, Z. F. Phys. **37**, 895 (1926)
2. W. Gordon, Z. F. Phys. **40**, 117 (1926)
3. P.A.M. Dirac, Proc. Roy. Soc. **A117**, 778 (1928)
4. L.C. Biedenharn, Found. Phys. **13**(1), 13 (1983)
5. W. Gordon, Z. F. Phys. **48**, 11 (1928)
6. C.G. Darwin, Proc. Roy. Soc. London **A118**, 654 (1928)
7. P.A.M. Dirac, Proc. Roy. Soc. **A126**, 801 (1930)
8. R. Oppenheimer, Phys. Rev. **35**, 562 (1930)
9. H. Weyl, *Gruppentheorie und Quantummechanik*, 2nd ed. (1930)
10. W. Pauli, Weiss, Helv. Phys. Acta **7**, 709 (1934)
11. W. Pauli, W. Heisenberg, Z. F. Phys. **56**, 1 (1929); Z. F. Phys. **59**, 168 (1930)
12. P.A.M. Dirac, Proc. Roy. Soc. **A112**, 661 (1926)
13. P.A.M. Dirac, Proc. Roy. Soc. **A114**, 243 (1927)
14. V.A. Fock, C.R. Leningrad, p. 267 (1933)
15. W. Furry, R. Oppenheimer, Phys. Rev. **45**, 245 (1934)
16. R. Oppenheimer, Phys. Rev. **35**, 461 (1930)
17. S. Weinberg, *The Quantum Theory of Fields-I* (Cambridge University Press)
18. V. Weisskopf, Phys. Rev. **56**, 72 (1939)
19. E.A. Uehling, Phys. Rev. **48**, 55 (1935)
20. H. Euler, B. Kochel, Naturwissenschaften **23**, 246 (1935)

21. H. Euler, W. Heisenberg, *Naturwissenschaften* **17**, 1 (1938)
22. N. Kemmer, V. Weisskopf, *Nature* **137**, 659 (1936)
23. W.E. Lamb, R.C. Retherford, *Phys. Rev.* **72**, 241 (1947)
24. H.M. Foley, P. Kusch, *Phys. Rev.* **73**, 412 (1948)
25. S. Pasternack, *Phys. Rev.* **54**, 1113 (1938)
26. S.S. Schweber, *QED and the Men Who Made It* (Princeton University Press, 1994)
27. J. Schwinger (ed.), *Quantum Electrodynamics* (Dover Publications, 1958)
28. F.J. Dyson, *Phys. Rev.* **75**, 486 (1949)
29. F.J. Dyson, *Phys. Rev.* **75**, 1736 (1949)
30. H.A. Bethe, *Phys. Rev.* **72**, 339 (1947)
31. H.A. Kramers, *Nuovo Cimento* **15**, 108 (1938)
32. H.A. Kramers, N.T. Natwrik, vol. 11 p. 134 (1944)
33. G. Wentzel, *Quantum Theory of Fields* (Interscience, NY, 1949)
34. V. Weisskopf, *Z. F. Phys.* **89**, 27 (1934)
35. V. Weisskopf, *Phys. Rev.* **56**, 72 (1939)
36. F.J. Dyson, *Phys. Rev.* **73**, 617 (1948)
37. G.E. Breit, *Phys. Rev.* **71**, 984 (1947)
38. J.E. Nafe, E.B. Nelson, I.I. Rabi, *Phys. Rev.* **71**, 914 (1947)
39. B. French, *The Shift of Energy Levels due to Radiative Coupling*, Ph.D. thesis, MIT (1948)
40. B. French, V. Weisskopf, *Phys. Rev.* **75**, 1240 (1948)
41. N.M. Kroll, W.E. Lamb, *Phys. Rev.* **75**, 388 (1949)
42. J. Schwinger, Two shakers of physics. *Lect. Notes. Phys.* **746**, 27–42 (2008)
43. P.A.M. Dirac, *Proc. Roy. Soc. London* **A316**, 453 (1932)
44. F. Bloch, K. Nordsieck, *Phys. Rev.* **52**, 54 (1937)
45. W. Pauli, Fierz, *Nuovo Cimento* **15**(3), 1 (1938)
46. S. Tomonaga, *Prog. Theor. Phys.* **1**, 27 (1946)
47. C.N. Yang, D. Feldman, *Phys. Rev.* **79**, 792 (1950)
48. R. Fukuda, M. Miyamoto, S. Tomonaga, *Prog. Theor. Phys.* **4**(2), 47 (1949)
49. R. Fukuda, M. Miyamoto, S. Tomonaga, *Prog. Theor. Phys.* **4**(2), 121 (1949)
50. S. Tomonaga, *Phys. Rev.* **74**, 224 (1948)
51. J. Schwinger, *Phys. Rev.* **73**, 416 (1948)
52. J. Schwinger, *Phys. Rev.* **75**, 898 (1949)
53. W. Pauli, F. Villars, *Rev. Mod. Phys.* **21**, 434 (1949)
54. J. Schwinger, *Phys. Rev.* **76**, 790 (1949)
55. J. Schwinger, *Phys. Rev.* **74**, 1212 (1948); *Phys. Rev.* **74**, 1439 (1948)
56. J. Schwinger, *Phys. Rev.* **75**, 651 (1949)
57. P.A.M. Dirac, *Phys. Zeits. Sowjetunion* **3**, 64 (1933)
58. N.D. Hari Dass, *Dirac and the Path Integral*. [arXiv:2003.12683](https://arxiv.org/abs/2003.12683)
59. L.M. Brown (ed.), *Feynman's Thesis: A New Approach to Quantum Theory* (World Scientific Publishing)
60. R.P. Feynman, *Rev. Mod. Phys.* **20**, 267 (1948)
61. J. Schwinger, *Phys. Rev.* **82**, 914 (1951)
62. R.P. Feynman, *Phys. Rev.* **76**, 749 (1949)
63. R.P. Feynman, *Phys. Rev.* **76**, 769 (1949)
64. R.P. Feynman, *Phys. Rev.* **80**, 440 (1950)
65. R.P. Feynman, *Phys. Rev.* **74**, 939 (1948)
66. G.C. Wick, *Phys. Rev.* **80**, 268 (1950)
67. A. Salam, *Phys. Rev.* **82**, 217 (1951)
68. V. Singh, *Phys. Rev.* **127**, 632 (1962)
69. R.J. Eden, P.V. Landshoff, D.I. Olive, J.C. Polkinghorne, *The Analytic S-matrix* (Cambridge University Press, 1966)



6.1 QFT-A Particle Perspective

In this chapter, we give a quick and self-contained primer on RQFT, emphasizing those non-perturbative features that will enable a connection to be made between causality and analyticity. It has become customary to introduce quantum fields as straightforward quantum generalizations of classical fields. But there is also a way of thinking of quantum fields as a way of coding our empirical knowledge. The experiments one conducts are all with particles, not fields, so a formulation entirely in terms of particles and their measurements seems like a minimalist approach. In classical field theories like electromagnetism and gravitation, fields provide an attractive local description. At least in the case of electromagnetism, a field-less description is certainly possible though it requires a technically harder non-local, i.e. “action at a distance” formulation. There is nothing inconsistent or undesirable about it. Even in these celebrated field theories, direct measurement of fields never happens, and all observations and measurements are through the intermediary of matter.

Though these aspects are seldom stressed, there is much about the particle world that is amazing, to say the least. Atomism is the bedrock of our particle world. A striking feature of that is that all our elementary particles are *identical*. Since this approach to field theory is hardly expounded either in the current-day research, or pedagogy, we will start with it. Of course, at some point, the introduction of fields, as a derivative description, will be seen to greatly facilitate matters, and we will switch over to that mode. At that point, we will have more to say about various fundamental aspects of quantum fields and how they are, fundamentally and paradigmatically, different from classical fields.

There are a countless number of books on QFT, and even the very good ones are many. For our narrative, which is a discussion of the non-perturbative links between causality, unitarity, analyticity and QFT, the short, but extremely lucid book by Barton

[1] is an ideal one. It has a clear discussion of various subtle (but essential to S-matrix theory) aspects like the asymptotic condition, interpolating fields, requirements like vacuum, one-particle stability (hardly found in many modern books), etc. It is the only book that I have seen which proves the *retarded commutator representation* of the LSZ reduction and we will see that this is actually central to establishing analyticity in QFT without resorting to perturbative methods. Another invaluable source is Bogoliubov and Shirkov's encyclopaedic book on QFT [2]. One of the earliest books on QFT which inspired even pioneers is Wentzel's book [3]. Among the more modern books that are highly recommended are *Particles, Sources and Fields* by Schwinger [4] (he calls it a research document and a textbook), *Relativistic Quantum Fields* by Bjorken and Drell [5] and *The Quantum Theory of Fields* by Steven Weinberg [6].

When it comes to relativistic matters, each author has a different convention wrt space-time metric. One group adopts $(-+++)$ with the logic that in Galilean Physics, space being Euclidean had the metric $(++++)$ and that time, the new-comer, must follow the majority rule or the view that new physics should not change old conventions as far as possible. This is the convention followed by Schwinger and Weinberg. Both Bjorken and Drell, as well as Barton, follow the convention $(+---)$. The metric convention percolates into many other issues like whether the Dirac γ -matrices are Hermitean or anti-Hermitean, which can be chosen to be real-symmetric, etc. There is nothing to do except set up your own table of translations.

6.1.1 Notations and Conventions

We adopt the same notations and conventions as Barton. The metric is $g_{\mu\nu} = (+, -, -, -)$. Therefore, a four-vector A is *spacelike* if $A^2 < 0$, *timelike* if $A^2 > 0$, and *lightlike* if $A^2 = 0$. A free particle of *rest mass* m has a four-momentum p_μ with the invariant $p^2 = p_0^2 - \mathbf{p}^2 = m^2$. Therefore, p is timelike. For photons, it is lightlike. The four-dimensional Laplacian, $\square = g^{\mu\nu} \partial_\mu \partial_\nu = -p^2$. A very important function is the *step function* $\theta(x)$

$$\theta(x) = 1 \quad x > 0 \quad \theta(x) = 0 \quad x < 0 \quad (6.1)$$

6.2 Quantum Fields from Particles

In a typical experiment, a set of particles, initially free, are brought into interaction among themselves, and scatter into free particles again. The transition between the initial and final states is coded in the S-matrix of Heisenberg. Though this is the general picture, some explanations are necessary. What one means “initially” are at time scales into the past that are much larger than the *characteristic time scales* associated with the interactions. Admittedly, this is vague and is “taken care of” by setting the initial times to $t = -\infty$. Likewise, spatial separations much larger than

the characteristic sizes of the interaction region are deemed to be at special infinity. These mathematical idealizations should always be treated with extreme care.

A more non-trivial issue arises out of the picture of treating the initial and final states of particles as “free”, equivalently, non-interacting. This crucially depends on how fast interactions fall off with distance. With interactions like the coulomb or gravitation, the fall-off is not sufficient to ensure free states even very far away from interaction. Here too, Quantum Mechanics provides a more satisfactory picture; the Schrödinger equation for electron-proton scattering can be exactly solved and one finds phases growing logarithmically with distance that mark deviations from plane or spherical waves. This is the notorious *Infrared Problem*. But the main motivations in this book are to understand the S-matrix for strong interactions where all interactions are finite-range. Of course, nature does not separate the interactions this way, and this limitation should always be kept in mind. With these cautionary remarks, we now formulate the so-called *asymptotic conditions*, which Barton treats as fundamental axioms.

To make the discussion uncluttered, we shall consider all particles to be spin-0 and with no other quantum numbers like charge, etc. Generalizations to include the more realistic cases is straightforward. The asymptotic conditions define the so-called in and out states:

States in the distant past ($t \rightarrow -\infty$) with a given number of free particles form a complete set, called *in* states, and are denoted by

$$|k_1, k_2, \dots k_n(in)\rangle \quad (6.2)$$

Likewise, states in the distant future ($t \rightarrow \infty$) with a given number of free particles form a complete set and are called *out* states

$$|k_1, k_2, \dots k_n(out)\rangle \quad (6.3)$$

Even at the level of QM, free particle states with definite momenta pose technical difficulties. The plane wave solutions are not normalizable. In the olden days, this was sought to be handled with the help of the so-called *Box Normalization*. Three-dimensional space was considered to be enclosed in a huge volume V with wavefunctions satisfying appropriate boundary conditions. This made the momentum states discrete. While in non-relativistic QM this posed no problems, it lacks relativistic covariance. The alternative is to use the continuum normalization, but that has its own mathematical subtleties. We shall go along with Barton and use box normalization with the proviso that in the end the box volume will be taken to infinity in which limit, the identification

$$V^{-1} \sum_{\mathbf{k}} \rightarrow (2\pi)^{-3} \int d\mathbf{k} \quad (6.4)$$

is made. Those who dislike box normalization, or feel uncomfortable with its use may use normalizations of their choice. The major conclusions remain unchanged.

It is worth emphasizing that in *Field Theories on a Lattice* (lot more on that in the later parts of our book), it is the box normalization that is natural. With box normalization, and the discrete nature of the momentum spectrum, one particle states can be normalized using *Kronecker Delta* instead of *Dirac Delta*.

$$\langle \mathbf{k} | \mathbf{k}' \rangle = \delta_{\mathbf{k}, \mathbf{k}'} \quad (6.5)$$

Barton introduces the next fundamental axiom which may be called the *Axiom of the Vacuum State*. *There exists a state $|0\rangle$ called the vacuum state* which has the properties (I have rearranged Barton's list according to what I think are the most fundamental):

- It is *unique* and *normalizable*, i.e. $\langle 0|0\rangle = 1$.
- It is invariant under *Lorentz transformations* as well as under all internal symmetries including the discrete ones. More succinctly, it has zero eigenvalues for all dynamical variables.
- It is the lowest energy eigenstate.

All the properties of the free particle states can be very compactly codified by first introducing the so-called *annihilation operators* $a_{\mathbf{k}}$ for each *mode* $\mathbf{k}$ with

$$a_{\mathbf{k}}|0\rangle = 0 \quad (6.6)$$

along with their algebra

$$[a_{\mathbf{k}}, a_{\mathbf{k}'}] = 0 \quad [a_{\mathbf{k}}, a_{\mathbf{k}'}^\dagger] = \delta_{\mathbf{k}\mathbf{k}'} \quad (6.7)$$

The n-particle states are now represented by

$$|k_1, k_2, \dots k_n\rangle = a_{\mathbf{k}_1}^\dagger a_{\mathbf{k}_2}^\dagger \dots a_{\mathbf{k}_n}^\dagger |0\rangle \quad (6.8)$$

These states need to be properly normalized. It is also useful to introduce the “number operator”

$$N_{\mathbf{k}} = a_{\mathbf{k}}^\dagger a_{\mathbf{k}} \quad (6.9)$$

By virtue of the algebra of the operators, it is easy to see that the eigenvalues of the number operators are integers, and they just count the number of particles of a given mode.

The asymptotic axiom introduces two completely different types of particle states, i.e. the *in* and the *out* types. One has to introduce two distinct types of $a, a^\dagger$ operators each obeying an algebra identical to what was introduced earlier. This is easily done by introducing $a^{(in)}, a^{(out)}$. The n-particle in and out states can be constructed with these. The S-matrix, which connects n-particle in-states to n-particle out-states, acts

on the vacuum state and the annihilation operators (consequently on the creation operators too) according to

$$S|0, in\rangle = |0, out\rangle \quad S a_{\mathbf{k}}^{(in)} S^\dagger = a_{\mathbf{k}}^{(out)} \quad S S^\dagger = S^\dagger S = 1 \quad (6.10)$$

This has to hold for each $\mathbf{k}$. So far, things have been rather general. Two additional axioms are introduced at this stage which will prove essential later on.

Vacuum Stability Axiom:

$$|0, in\rangle = |0, out\rangle = |0\rangle \quad (6.11)$$

The meaning of this axiom is that the state of nothingness, the vacuum, shall remain so even if the interactions are there, in the precise sense that $S \neq 1$. Yet, as Hawking has shown in his revolutionary work on Hawking Radiation, in Black Hole spacetimes, an in-vacuum can be transformed into an out-state with many particles. The vacuum stability condition is violated. How exactly this comes about is something worthwhile to understand.

One particle Stability:

$$|k, in\rangle = |k, out\rangle = |k\rangle \quad (6.12)$$

Why should one impose such an axiom, and what is its physical meaning? Its necessity arises from our wanting to derive the so-called LSZ reduction in the “retarded commutator form”. That may seem like an opaque technical reason, but its importance will become clearer later on. The physical meaning of this axiom is, however, much clearer. It means that even single particle states remain unchanged under time evolution. In other words, all single-particle states have to be stable and not decay. The unstable particles will then have to be understood in terms of ultimately decaying into the stable particles of the system, and S-matrix elements can only be determined when the in- and out-states are the stable particles. This is quite a severe restriction. For example, S-matrix for weak interactions could not have been strictly constructed. An incoming neutron state would end up as outgoing proton, electron and neutrino states, thereby violating this axiom. Once again, for the purely strongly interacting sector, no difficulties arise. It is interesting to speculate what would have happened if the weak interactions were also strong. Another point to emphasize is that even with an S-matrix for stable particles only, unstable particles can still show up as poles at complex masses.

Assembling the Quantum Field:

An in-quantum field can be constructed as follows:

$$\phi_{in}(x) = (2\pi)^{-3/2} \int \frac{d\mathbf{k}}{\sqrt{2k_0}} \{e^{-ik \cdot x} a_{\mathbf{k}}^{(in)\dagger} + e^{ik \cdot x} a_{\mathbf{k}}^{(in)}\} \quad (6.13)$$

with $k_0 = \sqrt{\mathbf{k}^2 + m^2}$. It is obviously an *operator valued field*. The out-field can likewise be constructed. The normalization factors have been so chosen as to agree

with the so-called *mode expansion* of free neutral scalar fields, found in textbooks. But we did not go through Lagrangeans, Hamiltonians, equal time commutation relations, nor what is called *quantization*.

It should be equally well emphasized that at this level of free particles, nothing was really gained by this exercise of assembling a field. But we will see the utility of this construction in going beyond the free behaviour. A natural question to ask, both from mathematical and physical points of view, is whether the free fields $\phi_{in}(x)(t \rightarrow -\infty)$ and $\phi_{out}(x)(t \rightarrow \infty)$ can be thought of as the past and future limits of an appropriate *interpolating field*. In between the distant past and the distant future, there are of course interactions and any such interpolating field will necessarily be an *interacting quantum field*.

Before we address this, when we assembled the quantum field from the $a_{\mathbf{k}}, a_{\mathbf{k}}^\dagger$, we quietly chose a *field normalization*. We could have chosen the field to be any constant multiple of what was chosen in Eq. (6.13). This is a matter of convenience which has no physical consequences. Once this normalization for the in and out fields is chosen, the normalization of the interpolating field can not be chosen at will. It has to be consistent with other normalizations that are inherent to QFT's; for example, the way *Equal Time Commutation Relations* (ETCR) are normalized. We will try to clarify this subtle point as we go along. So, a naive expectation for the interpolating field $\phi(x)$ such that

$$\phi(x) \rightarrow \phi_{in}(x)(t \rightarrow -\infty) \quad \phi(x) \rightarrow \phi_{out}(x)(t \rightarrow -\infty) \quad (6.14)$$

is wrong on several counts. We first state the correct asymptotic condition for the interpolating field $\phi(x)$ as given by Lehmann, Symanzik and Zimmermann [7]:

$$\langle \alpha | \phi^f(t) | \beta \rangle \xrightarrow[t \rightarrow -\infty]{} \sqrt{Z} \alpha | \phi_{in}^f(x) | \beta \rangle \quad (6.15)$$

and likewise for the out-fields. Here, ϕ^f stands for the *smeared quantum field* given by

$$\phi^f(t) = i \int d^3x f^*(\mathbf{x}, t) \overleftrightarrow{(\partial_0)} \phi(\mathbf{x}, t) \quad (6.16)$$

with $f(\mathbf{x}, t)$ being any normalizable solution of the Klein-Gordon equation:

$$(\square + m^2) f(x) = 0 \quad (6.17)$$

Barton, however, defines smearing simply by

$$\phi_{Barton}^f = \int d\mathbf{x} f(\mathbf{x}) \phi(x) \quad (6.18)$$

where $f(\mathbf{x})$ are normalizable functions. Without smearing the quantum fields are too *singular* in the sense they can excite *infinite norm* excitations from the vacuum. We

shall not explicitly use smearing anymore. We shall adopt the simpler asymptotic condition

$$\langle \alpha | \phi(x) | \beta \rangle \xrightarrow{x_0 \rightarrow -\infty} \sqrt{Z} \langle \alpha | \phi_{in}^f(x) | \beta \rangle \quad (6.19)$$

The LSZ-asymptotic condition, over and above such subtleties as smearing, has another very deep message about interpolating fields and that is that interpolation is only meaningful at the level of matrix elements, and not at an operator level. In quantum mechanics, this was never an issue, and it may seem counter-intuitive. But we will see why this is essential in QFT when we discuss the Kallen-Lehmann representation for two-point functions in Chap. 7. In Eq. (6.15), Z is an indication that the interpolating field and the asymptotic fields need not, in principle, be normalized the same way. Both Barton [1] and Haag [8] choose the convention that the matrix elements between vacuum and single-particle states, of both the in-out fields and interpolating fields, are *equal*, which is the case when $Z = 1$. Bjorken and Drell [5] and Weinberg [6], on the other hand, keep $Z \neq 1$. “Physics” of course can not depend on conventions. We shall be commenting on this important issue as we go along. To facilitate the comparison, we shall denote by $\phi_{BR}(x)$ the interpolating field with Barton’s normalization, and, by $\phi_{BJ}(x)$, the one with the Bjorken-Drell normalization. Summarizing

$$\phi_{BJ}(x) = \sqrt{Z} \phi_{BR}(x) \quad (6.20)$$

In Eq. (6.13), the “modes” had been taken to be plane waves which are normalizable only in a box. It is desirable to consider also the complete set of positive-frequency normalizable (box or otherwise) solutions of the Klein-Gordon equation:

$$(\square + m^2) f_\alpha(x) = 0 \quad i \int d\mathbf{x} f_\alpha^*(x) \left(\overleftrightarrow{\frac{\partial}{\partial x_0}} \right) f_\beta(x) = \delta_{\alpha\beta} \quad (6.21)$$

If box normalization is not used these can be considered to be wave packets closely approximating plane waves. A new in-field can be assembled with the help of $f_\alpha(x)$, and the creation-annihilation operators labelled accordingly. For all practical purposes, f_α can just be taken to be plane waves. As expressed earlier, we shall leave the f_α as general as possible. The modified relations are gathered below:

$$\phi_{in}(x) = \sum_{\alpha} \{ a_{\alpha}^{(in)} f_{\alpha}(x) + a_{\alpha}^{(in)\dagger} f_{\alpha}^*(x) \} \quad (6.22)$$

$$[a_{\alpha}, a_{\alpha'}^{\dagger}] = \delta_{\alpha\alpha'} \quad (6.23)$$

A relation that will be frequently used is

$$a_{\alpha}^{(in)} = i \int d\mathbf{x} f_{\alpha}^* \left(\overleftrightarrow{\partial_0} \right) \phi_{in}(x) \quad (6.24)$$

Likewise, for the out-fields. An important quantity in QFT is

$$i\Delta(x - x') = [\phi_{in}(x), \phi_{in}(x')] \quad (6.25)$$

From the algebra of Eq.(6.23), it is clear that $\Delta(x - x')$ is a number, and not an operator. To see that it only depends on $(x - x')$, let us introduce the four-momentum operator, first for the particle description, and then for the constructed in-out fields.

At the free particle level, $|k\rangle$ is the eigenstate of the momentum Operator P_μ with eigenvalue k_μ . It is straightforward to construct this operator

$$P_\mu = \sum_{\mathbf{k}} k_\mu a_{\mathbf{k}}^\dagger a_{\mathbf{k}} \quad (6.26)$$

This works for both the in and out particle states. It is straightforward to see that (on noting that the $a, a^\dagger$ operators above are labelled by momentum)

$$[P_\mu, \phi_{in}(x)] = i \partial_\mu \phi_{in}(x) \quad (6.27)$$

Equivalently,

$$e^{-iP \cdot a} \phi_{in}(x) e^{iP \cdot a} = \phi_{in}(x + a) \quad (6.28)$$

In other words, at the field level, we discover the alternate meaning of the momentum operator as a *displacement* or *translation* operator. While the concept of momentum eigenstates ceases to be useful for interacting particles, the concept of a displacement operator continues to be meaningful even for interacting fields. Therefore, we extend the operation of P_μ for the interpolating fields also and write

$$[P_\mu, \phi(x)] = i \partial_\mu \phi(x) \quad e^{-iP \cdot a} \phi(x) e^{iP \cdot a} = \phi(x + a) \quad (6.29)$$

To prove that $\Delta(x - y)$ as defined above is indeed a function only of $x - y$, first one notes that since it is a number, it equals its vacuum expectation value, and then use Eq.(6.28):

$$\begin{aligned} \langle 0 | [\phi_{in}(x), \phi_{in}(y)] | 0 \rangle &= \langle 0 | e^{-P \cdot x} \phi_{in}(0) e^{iP \cdot x} e^{-P \cdot y} \phi_{in}(0) e^{iP \cdot y} | 0 \rangle - x \leftrightarrow y \\ &= \langle 0 | \phi_{in}(0) e^{-P \cdot (x-y)} \phi_{in}(0) | 0 \rangle - x \leftrightarrow y \end{aligned} \quad (6.30)$$

which depends on x, y only through the combination $(x - y)$.

The dual function of P_μ at both the particle and the field levels is noteworthy. Barton uses these properties of the momentum operator to derive two important and rather counter-intuitive results. He considers the matrix elements $\langle 0 | \phi(x) | k \rangle$, of the interpolating field between the vacuum state $|0\rangle$, and the one-particle state $|k\rangle$. By the one-particle stability axiom, $|k\rangle$ is both the in-state and an out-state. It is only the $\phi_{in,out}$ that satisfies the Klein-Gordon equation. Nevertheless,

$$\square \langle 0 | \phi(x) | k \rangle = -\langle 0 | [P_\mu, [P_\mu, \phi(x)] | k \rangle = -m^2 \langle 0 | \phi(x) | k \rangle \quad (6.31)$$

In other words, both the matrix elements $\langle 0|\phi(x)|k\rangle$, for the in- and out-states, are solutions of the **free** Klein-Gordon equation despite the interpolating field $\phi(x)$ not satisfying the Klein-Gordon equation. Both of them are also momentum eigenstates with the same eigenvalue k_μ . This can be seen as follows:

$$i\partial_\mu \langle 0|\phi(x)|k\rangle = \langle 0|[P_\mu, \phi(x)]|k\rangle = k_\mu \langle 0|\phi(x)|k\rangle \quad (6.32)$$

Combining this with the asymptotic condition of Eq. (6.19), one gets

$$\langle 0|\phi_{BJ}(x)|k\rangle = \sqrt{Z} f_k(x) \quad \langle 0|\phi_{BR}(x)|k\rangle = f_k(x) \quad (6.33)$$

a rather surprising result. As already commented, Barton's choices for field normalizations are such that $Z = 1$ in all his formulae, so far. Barton recasts this in another very useful form by introducing the notion of a *source* $\eta(x)$ defined by

$$\eta_{BR}(x) \equiv (\square + m^2) \phi_{BR}(x) \quad (6.34)$$

This does not assume knowledge of any equation satisfied by $\phi(x)$ itself. It is just a definition of the source term. It then follows from Eq. (6.31) that (for both BR and BJ)

$$\langle 0|\eta(x)|k\rangle = 0 \quad (6.35)$$

We now turn to the important issue of the commutator $[\phi(x), \phi(x')]$ of the interpolating fields. Unlike the function $\Delta(x - y)$, the commutator of the free asymptotic fields, the commutator of the interpolating fields is not a number, but an operator. The structure of the entire commutator will obviously depend on a lot of details about interactions. What can be said is about it when x, y are *space-like* separated; microcausality, that we discussed previously tells us that the commutator will vanish then. What about the situations when x, y are not space-like separated, i.e. time-like or light-like separated? Instead of trying to answer in full generality, let us focus on the vacuum expectation value of this commutator

$$\Delta'(x - y) \equiv \langle 0|[\phi(x), \phi(y)]|0\rangle \quad (6.36)$$

That this function also depends only on $x - y$ follows from the same line of reasoning we used for $\Delta(x - y)$ i.e. it is just a consequence of translational invariance. Then, microcausality dictates that $\Delta'(x - y)$ should also vanish when $x - y$ is space-like. We have so far not introduced any Lagrangean. If the interactions are not of the *derivative type*, as in QED, to be later introduced QCD, and the so-called $\lambda \phi^4$ interaction for scalar fields, and, if we momentarily give up insisting on manifest covariance, the ETCR for these cases would take the form (modulo the specific field normalizations used)

$$[\phi_{BJ}(\mathbf{x}, t), \dot{\phi}_{BJ}(\mathbf{y}, t)] = i\delta(\mathbf{x} - \mathbf{y}) \quad (6.37)$$

Once again, the issue of normalizations in QFT props up. This form of the ETCR, adopted by [5,6] is only consistent when the choice $Z \neq 1$ is made. On the other hand, if the choice made is $Z = 1$ [1,8], the ETCR (even in the restricted circumstances of non-derivative couplings) has to be modified (see Eq. (8.6) of [1]):

$$[\phi_{BR}(\mathbf{x}, t), \dot{\phi}_{BR}(\mathbf{y}, t)] = -i Z_3^{-1} \delta(\mathbf{x} - \mathbf{y}) \quad (6.38)$$

Clearly, the physics of Z_3 of one convention, and of Z in another are related. In fact, we shall argue that numerically $Z_3 = Z$. This would imply, for $\Delta'(x - y)$ (for $Z \neq 1$)

$$\frac{\partial}{\partial y_0} \Delta'_{BJ}(x - y)|_{y_0=x_0} = i \delta(\mathbf{x} - \mathbf{y}); \quad \frac{\partial}{\partial y_0} \Delta'_{BR}(x - y)|_{y_0=x_0} = i Z_3^{-1} \delta(\mathbf{x} - \mathbf{y}) \quad (6.39)$$

In the next chapter, we take a much more detailed look at these issues by non-perturbative analysis. That will also explain why the naive or the so-called strong asymptotic conditions lead to contradictions.

References

1. G. Barton, Introduction to advanced field theory. Intersci. Tracts Phys. Astron. **22** (1963)
2. N.N. Bogoliubov, D.V. Shirkov, *Introduction to the Theory of Quantised Fields* (Interscience, NY, 1959)
3. G. Wentzel, *Quantum Theory of Fields* (Interscience, NY, 1949)
4. J. Schwinger, *Particles, Sources, and Fields*, vol. I (Addison-Wesley Publishing Company, 1970)
5. J.D. Bjorken, S. Drell, *Relativistic Quantum Fields* (McGraw Hill Publishers)
6. S. Weinberg, *The Quantum Theory of Fields -I* (Cambridge University Press)
7. H. Lehmann, K. Symanzik, W. Zimmermann, *Nuovo Cimento* **1**, 1425 (1955)
8. R. Haag, *Local Quantum Physics*, 2nd ed. (Springer Publications, 1991)

The Kallen-Lehmann Representation

7

7.1 Introduction

In this chapter, we shall discuss how important results can be obtained in QFT without resorting to perturbative methods as done in QED. There is a vast literature on this subject but we shall be content with showing how such methods can be applied to two important problems. The first is a non-perturbative analysis of the so-called two-point function $\Delta'(x - y)$ introduced in Chap. 5. To illustrate the essential concepts and techniques, we shall continue to work with spinless fields. Generalization to realistic fields like Dirac and Maxwell fields can be found in the books on field theory listed earlier. The second is the so-called Lehmann-Symanzik-Zimmermann (LSZ) reduction formulae. This relates the S-matrix elements to Fourier transforms of either the so-called *time ordered products* (we already encountered these while discussing Dyson's works in QED) or to the Fourier transforms of the so-called *retarded commutators*. In this latter form, they provide a crucial link between analyticity and microcausality in QFT, in a way very close to the spirit as well as the details of the Kramers-Kronig dispersion relations.

The pioneering breakthroughs in this line of thinking were done by Kallen, Lehmann, Jost, Symanzik, Zimmermann, Dyson, Bogoliubov, etc. The techniques to be used for both the results are what were described in Chap. 6, and involve in an essential way the asymptotic in and out fields, the interpolating fields and the asymptotic conditions. Towards this end, we continue with Barton's exposition as it is lucid and uncluttered.

7.2 The Kallen-Lehmann Representation

The object of our interest is (unless clearly specified, the interpolating field $\phi(x)$ can be taken to be with either BJ or BR normalization)

$$\Delta'(x-y) \equiv \langle 0 | [\phi(x), \phi(y)] | 0 \rangle \quad (7.1)$$

Following Barton, we introduce another object, closely related to the above:

$$\Delta^{(+)'}(x-y) \equiv \langle 0 | \phi(x) \phi(y) | 0 \rangle \quad (7.2)$$

That this too is only a function of $(x-y)$ follows from the same arguments of translation invariance that were used for $\Delta(x-y)$ and $\Delta'(x-y)$ in Chap. 5. We shall make it explicit now on using

$$\phi(x) = e^{-iP \cdot x} \phi(0) e^{iP \cdot x} \quad (7.3)$$

Introducing a complete set of states, we rewrite Eq. (7.2)

$$\begin{aligned} \Delta^{(+)'}(x-y) &= \sum_s \langle 0 | \phi(x) | s \rangle \langle s | \phi(y) | 0 \rangle \\ &= \sum_s |\langle 0 | \phi(0) | s \rangle|^2 e^{ip_s \cdot (x-y)} \end{aligned} \quad (7.4)$$

It should be noted that $|s\rangle$ includes arbitrary *multiparticle* states and p_s is their total momentum. Additionally, p_s is time-like and $p_{s,0}$ is always *positive*. The trick now is to use the identity

$$\int d^4p \delta^{(4)}(p_s - p) \theta(p_{s,0}) = 1 \quad (7.5)$$

Introducing this in Eq. (7.4) and interchanging the summation over s and integration over p , one obtains

$$\Delta^{(+)'}(x-y) = (2\pi)^{-3} \int d^4p e^{ip \cdot (x-y)} \theta(p_0) \rho(p^2) \quad (7.6)$$

where

$$\rho(p^2) = (2\pi)^3 \sum_s \delta^{(4)}(p_s - p) \theta(p) |\langle 0 | \phi(0) | s \rangle|^2 \quad (7.7)$$

Obviously, $\rho(p^2)$ for the BR and BJ fields differ by a proportionality constant. That $\rho(p^2)$ is a function of p^2 alone follows from Lorentz invariance. Writing $\rho(p^2)$ as

$$\rho(p^2) = \int_0^\infty da^2 \delta(p^2 - a^2) \rho(a^2) \quad (7.8)$$

It is important to note that since p^2 here is time-like, $a^2 \geq 0$. One further rewrites the expression for $\Delta^{(+)'}(x)$ as

$$\Delta^{(+)'}(x) = \int_0^\infty da^2 \rho(a^2) \int d^4 p (2\pi)^{-3} \theta(p) \delta(p^2 - a^2) e^{ip \cdot x} \quad (7.9)$$

We have interchanged summations and integrations rather liberally, and one will have to justify these. Such issues are clarified in the original works [1, 2]. For the moment let us go on. This is a remarkable representation (also called *spectral representation*) for the exact non-perturbative two-point function and hardly any dynamical inputs were invoked. To appreciate the meaning and significance of this representation note that

$$\Delta^{(+)}(x|a^2) \equiv \langle 0|\phi_{in}(x|a^2)\phi_{in}(0|a^2)|0\rangle = \int d^4 p (2\pi)^{-3} \theta(p) \delta(p^2 - a^2) e^{ip \cdot x} \quad (7.10)$$

for a free field of mass $m^2 = a^2$. Summarizing, what Kallen [1] (for QED) and Lehmann [2] showed was

$$\Delta^{(+)'}(x) = \int da^2 \rho(a^2) \Delta^{(+)}(x|a^2) \quad (7.11)$$

We remind the reader once again that $\Delta^{(+)}$ are for free fields and $\Delta^{+'}$ for interacting fields. There is a plethora of such two-point functions (see Appendix C of [3]), and for all of them similar representations can be worked out along the same lines. Take, for example, the Feynman propagation function defined by

$$i\Delta_F(x-y) \equiv \langle 0|T\{\phi(x)\phi(y)\}|0\rangle \quad (7.12)$$

The spectral representation for this in the momentum representation is

$$\Delta'_F(p) = \int_0^\infty da^2 \frac{\rho(a^2)}{p^2 - a^2 + i\epsilon} \quad (7.13)$$

Now, we address several important issues in the derivation of the Kallen-Lehmann representation.

The first issue is the cavalier manner in which sums and integrals were interchanged, and even among integrations, the way the order of integrations was interchanged. Those can be justified only if the resulting expressions are convergent. It is quite clear from the expressions that the spectral representations do not always exist (converge). That crucially depends on how fast the spectral functions $\rho(a^2)$ grow with a^2 . When the growth is such that the integrals exist, the spectral representation is of the *unsubtracted* form. If, on the other hand, the unsubtracted form does not converge, one may need one or more “subtractions”. For example, the once subtracted form would be

$$\Delta'_F(p^2) = \Delta'_F(q^2) + (p^2 - q^2) \int_0^\infty da^2 \frac{\rho(a^2)}{(a^2 - q^2)(p^2 - a^2 + i\epsilon)} \quad (7.14)$$

In such a case, the value of $\Delta'_F(q^2)$ at some arbitrarily chosen q^2 must be specified. Exactly similar issues and similar terminology will be encountered in the context of *Dispersion Relations*. It is worthwhile to view this issue also in terms of the large p^2 -asymptotics. For free (in-out) fields, it is obvious that

$$\Delta_F(p) = \frac{1}{p^2 - m^2 + i\epsilon} \xrightarrow{p^2 \rightarrow \infty} \frac{1}{p^2} \quad (7.15)$$

In the unsubtracted case for the full two-point function, this asymptotic behaviour is the same, modulo a constant factor

$$\Delta'_F(p^2) \xrightarrow{p^2 \rightarrow \infty} \frac{\int_0^\infty da^2 \rho(a^2)}{p^2} \quad (7.16)$$

In the once-subtracted example of Eq. (7.14), this asymptotic behaviour is dramatically changed:

$$\Delta'_F(p^2) \xrightarrow{p^2 \rightarrow \infty} \int_0^\infty da^2 \frac{\rho(a^2)}{(a^2 - q^2)} \quad (7.17)$$

The reader is also referred to Weinberg (footnote on p. 460 [4]) for remarks of a different nature in this connection. In the context of the issue of divergences in field theories, it would be of interest to have propagators fall off even faster than the free ones. But the Kallen-Lehmann representation shows that as long as the spectral function is positive, this can never happen. In that sense, it is remarkable that String Theory produces exponentially falling two-point functions.

From the expression Eq. (7.7), it readily follows that $\rho(p^2)$ is always *positive* when p^2 is time-like. It vanishes when p^2 is space-like. Some explanation is necessary at this point. We have tacitly assumed that the stable particles have positive energy and that $m^2 > 0$ for them. In other words, *Tachyons* are not considered to be among our “particles”. It is further assumed that m^2 for single particles assumes only discrete values. Stable-bound states are another issue. All such considerations go under the name of *Spectrum Conditions* on P_μ (see [3] Sect. 16.2).

It turns out that the contribution of the one-particle states to $\rho(a^2)$ can be exactly determined, thanks to the important result Eq. (6.33) of Chap. 6, based on the stability axioms. We write this result here explicitly owing to its importance (separately for the BR and BJ cases):

$$\langle 0 | \phi_{BR}(x) | k \rangle = \langle 0 | \phi_{in}(x) | k \rangle \quad \langle 0 | \phi_{BJ}(x) | k \rangle = \sqrt{Z} \langle 0 | \phi_{in}(x) | k \rangle \quad (7.18)$$

An equivalent relation for out-fields also holds. Substituting this in the one-particle contribution to $\rho(p^2)$, it immediately follows that the one-particle contribution to $\rho(p^2)$ is

$$\rho_{BJ}^{sing}(a^2) = Z \delta(a^2 - m^2) \quad \rho_{BR}^{sing}(a^2) = \delta(a^2 - m^2) \quad (7.19)$$

In accordance with the assumed spectrum conditions on P_μ , the single particle contribution is an *isolated pole* in a^2 -plane. Because of the simple structure of $\rho(p^2)$, multi-particle contributions can likewise be identified. Barring bound states, a n-particle contribution starts at a threshold of $a_n^2 = (nm)^2$, and the contributions from different n-values are additive. Denoting the n-particle contribution by $\sigma_n(a^2)$, the entire multiparticle contribution by $\sigma(a^2)$, one can write

$$\sigma(a^2) = \sum_{n=2} \theta(a^2 - n^2 m^2) \sigma_n(a^2) \quad (7.20)$$

In summary,

$$\begin{aligned} \rho_{BJ}(a^2) &= Z \delta(a^2 - m^2) + \theta(a^2 - 4m^2) \sigma_{BJ}(a^2) \\ \rho_{BR}(a^2) &= \delta(a^2 - m^2) + \theta(a^2 - 4m^2) \sigma_{BR}(a^2) \end{aligned} \quad (7.21)$$

Assuming no subtractions, we can present the expression for Δ'_F , in momentum space, as

$$\Delta'_{F,BJ}(p^2) = \frac{Z}{p^2 - m^2 + i\epsilon} + \int_{4m^2}^{\infty} da^2 \sigma_{BJ}(a^2) \frac{1}{p^2 - a^2 + i\epsilon} \quad (7.22)$$

and,

$$\Delta'_{F,BR}(p^2) = \frac{1}{p^2 - m^2 + i\epsilon} + \int_{4m^2}^{\infty} da^2 \sigma_{BR}(a^2) \frac{1}{p^2 - a^2 + i\epsilon} \quad (7.23)$$

It is actually better to present this as

$$\Delta'_{F,BJ}(p^2) = \int_0^{\infty} da^2 \frac{Z \delta(a^2 - m^2) + \theta(a^2 - 4m^2) \sigma_{BJ}(a^2)}{p^2 - a^2 + i\epsilon} \quad (7.24)$$

and,

$$\Delta'_{F,BR}(p^2) = \int_0^{\infty} da^2 \frac{\delta(a^2 - m^2) + \theta(a^2 - 4m^2) \sigma_{BR}(a^2)}{p^2 - a^2 + i\epsilon} \quad (7.25)$$

The corresponding expression for $\Delta'(x)$ in coordinate representation will also turn out to be important. We will display this both for Barton's(BR) choice of normalization $Z = 1$, as well as for Bjorken-Drell (BJ) normalization of $Z \neq 1$. This is to clearly bring out how such freedoms in the choice of normalizations do not change underlying physics.

$$\Delta'(x)_{BJ} = \int_0^{\infty} da^2 \{Z \delta(a^2 - m^2) + \theta(a^2 - 4m^2) \sigma_{BJ}(a^2)\} \Delta(x|a^2) \quad (7.26)$$

and,

$$\Delta'(x)_{BR} = \int_0^\infty da^2 \{ \delta(a^2 - m^2 + \theta(a^2 - 4m^2) \sigma_{BR}(a^2) \} \Delta(x|a^2) \quad (7.27)$$

where $\Delta(x|a^2)$ is the free particle $\Delta(x)$ for a particle of mass a^2 . We also separately display the vacuum expectation values of ETCR's for BR and BJ normalizations:

$$\frac{\partial}{\partial y_0} \Delta'_{BR}(x - y)|_{y_0=x_0} = i Z_3^{-1} \delta(\mathbf{x} - \mathbf{y}) \quad (7.28)$$

and,

$$\frac{\partial}{\partial y_0} \Delta'_{BJ}(x - y)|_{y_0=x_0} = i \delta(\mathbf{x} - \mathbf{y}) \quad (7.29)$$

Now, we turn to similar expressions for the free field case $\Delta(x - y)$ where, of course, there is no distinction between the BJ and BR normalizations. One has

$$\frac{\partial}{\partial y_0} \Delta(x - y|a^2)|_{y_0=x_0} = i \delta(\mathbf{x} - \mathbf{y}) \quad (7.30)$$

The remarkable feature of this is that it is independent of a^2 , reflecting the mass independence of the ETCR. Combining these, we get the conditions

$$Z_3^{-1} = 1 + \int_{4m^2}^\infty da^2 \sigma_{BR}(a^2) \quad (7.31)$$

for the BR normalization, and,

$$1 = Z + \int_{4m^2}^\infty da^2 \sigma_{BJ}(a^2) \quad (7.32)$$

On recalling the translation between BR and BJ fields as given in Eq. (6.20)

$$\phi_{BJ}(x) = \sqrt{Z} \phi_{BR}(x) \quad (7.33)$$

the $\sigma_{BJ}(a^2)$ and $\sigma_{BR}(a^2)$ should also be related as

$$\sigma_{BJ}(a^2) = Z \sigma_{BR}(a^2) \quad (7.34)$$

On using this, we see that Eqs. (7.31) and (7.32) become identical on identifying $Z_3 = Z$. Thus, these differing choices of field normalizations have the same physical consequences, which can be further explored with either of these choices. Let us do so with the BJ version. Before that, the following observation is extremely important.

In the BR normalization, the interacting ETCR differs from the free field ETCR by the factor Z_3^{-1} while the asymptotic condition relating the vacuum to single particle matrix elements has no such factors. In the BJ case, while the asymptotic condition

has an extra factor of $\sqrt{Z}$, the ETCR does not. Either way, there is always an extra factor in relating the interacting field expressions to the corresponding free field ones. This has the profound implication that there can be no unitary transformations relating free and interacting fields.

Now we end this discussion by looking at the possible extreme values of Z . Equation (7.32), in view of the positive nature of $\sigma(a^2)$, is tantamount to $0 \leq Z \leq 1$. Can Z ever take the extreme values of $(0, 1)$? Now, it is clear that Z can take the value 1 (in the BR case this amounts to $Z_3 = 1$) only when the multi-particle contribution is exactly 0, but that can only happen when the fields are free. In perturbative treatments of QED, because of the weakness of the coupling, Z is very close to 1 but even there it is not exactly equal to 1. This has some consequences for Dyson's proof of the equivalence between Tomonaga-Schwinger theories on the one hand and Feynman theory on the other. We had commented on this in Chap. 4 and had pointed out a missing link which would relate interacting and free fields. Dyson's proof requires this in operator form, or equivalently, a strong form of the asymptotic condition. But we have now shown that such a strong asymptotic condition contradicts the Kallen-Lehmann representation in that even in the joint limit of $x_0, y_0 \rightarrow \infty$ keeping $x_0 - y_0$ fixed, the $\Delta(x - y)$ and $\Delta'(x - y)$ can never be the same. In light of this, Dyson's proof needs a careful reappraisal.

Finally, what about $Z = 0$? This can happen when interactions become so strong that

$$\int_{4m^2}^{\infty} da^2 \sigma_{BJ}(a^2) = 1 \quad (7.35)$$

There is nothing in our considerations so far that can preclude such a possibility. In such a case, the single particle pole in the two-point function simply disappears. This can have profound consequences. For example, for the Higgs field this can open up the possibility that while the field can have non-vanishing vacuum expectation values, no particle can be associated with such a field.

7.2.1 Lessons for Analytic S-Matrix

Following Barton, one defines a function of a *complex variable* z by

$$\Delta'_F(z) = \frac{Z}{z - m^2} + \int_{4m^2}^{\infty} da^2 \sigma(a^2) \frac{1}{z - a^2} \quad (7.36)$$

Notice there is no need for the $i\epsilon$ term, as z is already a complex variable. As a function of the complex variable z , Δ'_F has an isolated pole at $z = m^2$, and a cut running from $z = 4m^2$ to $z = \infty$. Otherwise, Δ'_F is *analytic* in the entire z -plane. If multi-particle bound states exist, there would be additional poles, but the pole at $z = m^2$ is minimal. The “discontinuity” across the cut is $\pi \sigma(z)$.

The lesson from all this is that the physically relevant $\Delta'_F(p^2)$ for *real* values of z , can be thought of as the *boundary value* of a complex function which is analytic everywhere except for a few poles and a branch cut. It is also important to observe

that $\Delta'_F(z)$ is a *real analytic function*. Equation (7.36) is prototypical of the so-called *Dispersion Relations*, albeit for the two-point function. The aim of the analytic S-matrix program was that the scattering amplitudes likewise are also to be thought of as almost analytic functions (i.e. except for poles and cuts). It is important that they are also *real analytic functions*. The physical scattering amplitudes are to be thought of as boundary values of this analytic function as its complex arguments approach their real physical values. With scattering amplitudes, the number of complex variables to consider is more and the problem is considerably more difficult.

We shall have a lot more to say on this as we go along.

References

1. G. Kallen, *Helv. Phys. Acta* **25**, 417 (1954)
2. H. Lehmann, *Nuovo Cim.* **11**, 342 (1954)
3. J.D. Bjorken, S. Drell, *Relativistic Quantum Fields* (McGraw Hill Publishers)
4. S. Weinberg, *The Quantum Theory of Fields-I* (Cambridge University Press)

The Lehmann Symanzik Zimmermann (LSZ) Formalism

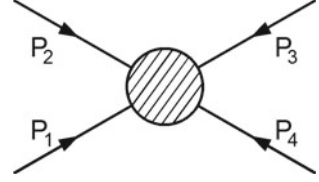
8

8.1 The LSZ Reduction Formulae

The second non-perturbative result we shall demonstrate is the so-called *reduction formulae* of Lehmann-Symanzik-Zimmermann (LSZ). This should not be confused with the *LSZ Theorem* which is a theorem about the asymptotic behaviour of the so-called *Vertex Function*. Again, Barton gives a very lucid account of the latter too. Interested readers should refer to Chap. 10 of his book. The LSZ theorem tells us, in a non-perturbative manner, the large momentum transfer behaviour of vertex functions. Interestingly, the Kallen-Lehmann representation plays a crucial role in its proof.

The LSZ reduction formulae, apart from giving a non-perturbative definition of S-matrix in RQFT, also provide a crucial link between a so-defined S-matrix and various aspects of analyticity that will be central to the considerations of our book. We shall continue to follow Barton's [1] treatment. The reader is highly recommended to follow the treatments in the book by Bjorken and Drell [2], as well as Gasirowicz's book [3] for alternate approaches. Here, we follow Barton's treatment for the case of neutral scalar particles. Though that is not the most realistic or physically the most interesting, it has the advantage of being the simplest, which enables the most salient aspects of the LSZ method to be grasped without the technical complications of spin, additional degrees of freedom, etc. Bjorken and Drell develop the LSZ reductions to all these physically relevant cases. Gasirowicz applies to the case of pion-nucleon scattering which brings out the subtleties of both fermion and bosonic fields, unequal mass scattering, etc. Some of the important intermediate steps of Gasirowicz do not work, for example in Barton's treatment, because both the initial and final states contain only one particle of each type.

From the point of view of this book, what is important is a treatment from which analyticity in QFT follows in a transparent manner. For that, a blend of Barton's treatment where all particles are spin-less (and identical), with Gasirowicz's treat-

Fig. 8.1 $2 \rightarrow 2$ scattering

ment without the spin complications but maintaining unequal masses will turn out to be optimal.

Though the method is applicable to scattering of an arbitrary number of particles, let us illustrate it with scattering of two incoming particles into two outgoing ones. We shall see that this is general enough to encompass all the essential features of the S-matrix. This process is symbolically represented in Fig. 8.1. In what follows, the one-particle stability condition which allows one particle in-states of given momentum to be identified with one particle out-states of the same momentum, will be heavily made use of. The reader is advised a lot of caution in this as a careless application can lead to completely wrong results. So, the starting point is

$$\mathcal{M} = \langle \alpha' \beta', out | \alpha \beta, in \rangle \equiv \langle \alpha' \beta', in | S | \alpha \beta, in \rangle \quad (8.1)$$

Because of the unitarity of the S-matrix,

$$S^{-1} = S^\dagger \quad (8.2)$$

the same matrix element can also be expressed in terms of the out-fields:

$$\mathcal{M} = \langle \alpha' \beta', out | \alpha \beta, in \rangle \equiv \langle \alpha' \beta', out | S^\dagger | \alpha \beta, out \rangle \quad (8.3)$$

We shall work with the in-basis. It is clear that

$$|\alpha \beta, in \rangle = a_\beta^{(in)\dagger} |\alpha \rangle \quad (8.4)$$

Obviously, this can be done for any of the four particles involved. Note that by writing $|\alpha, in \rangle$ as $|\alpha \rangle$, the one-particle stability has been made use of. Next step involves using the definition of $a_\beta^{(in)\dagger}$:

$$a_\beta^{(in)\dagger} = i \int d\mathbf{y} \phi_{in}(\mathbf{y}) \overleftrightarrow{\frac{\partial}{\partial y_0}} f_\beta(\mathbf{y}) \quad (8.5)$$

$f_\beta(\mathbf{y})$ being the normalized Klein-Gordon eigenfunctions introduced earlier. On using this,

$$\mathcal{M} = i \int d\mathbf{y} \langle \alpha' \beta', out | \phi_{in}(\mathbf{y}) | \alpha \rangle \overleftrightarrow{\frac{\partial}{\partial y_0}} f_\beta(\mathbf{y}) \quad (8.6)$$

Barton then uses the identity

$$\int d\mathbf{y}|_{y_0 \rightarrow -\infty} = \int d\mathbf{y}|_{y_0 \rightarrow \infty} - \int dy \frac{\partial}{\partial y_0} \quad (8.7)$$

and rewrites $\mathcal{M}$ as

$$\begin{aligned} \mathcal{M} = & -i \int dy \frac{\partial}{\partial y_0} \{ \langle \alpha' \beta', out | \phi(y) | \alpha \rangle \frac{\overleftrightarrow{\partial}}{\partial y_0} f_\beta(y) \} \\ & + i \int d\mathbf{y} \langle \alpha' \beta', out | \phi(y) | \alpha \rangle \frac{\overleftrightarrow{\partial}}{\partial y_0} f_\beta(y) |_{y_0 \rightarrow \infty} \end{aligned} \quad (8.8)$$

To evaluate the second term, we use the asymptotic condition, and the weaker form of it suffices as we are only dealing with matrix elements of $\phi(y)$. Consequently, the second term is

$$i \int d\mathbf{y} \langle \alpha' \beta', out | \phi_{out}(y) | \alpha \rangle \frac{\overleftrightarrow{\partial}}{\partial y_0} f_\beta(y) = \langle \alpha' \beta', out | a_\beta^{(out)\dagger} | \alpha \rangle \quad (8.9)$$

where use has again been made of the definition of the creation operator in terms of the field. It should however be carefully noted that we have passed from an in-creation operator to an out-creation operator in the process. Now, we again make use of the one-particle stability condition to write $|\alpha\rangle$ as $|\alpha, out\rangle$ and conclude that the second term is nothing but

$$\langle \alpha' \beta', out | \alpha \beta, out \rangle \quad (8.10)$$

This is just the matrix element of *identity*. It is interesting that the LSZ formalism naturally achieves the split $S = 1 + \bar{T}(\bar{T}$ differs from the conventionally defined T-matrix by an overall energy-momentum conservation delta-function) non-perturbatively. The rest of $\mathcal{M}$ is, then,

$$-i \int dy \{ \langle \alpha' \beta', out | \phi(y) | \alpha \rangle \frac{\partial^2 f_\beta}{\partial y_0^2} - (\frac{\partial^2}{\partial y_0^2} \langle \alpha' \beta', out | \phi(y) | \alpha \rangle) f_\beta \} \quad (8.11)$$

The $f_\beta(y)$ satisfies the free Klein-Gordon equation

$$\frac{\partial^2 f_\beta(y)}{\partial y_0^2} = (\nabla^2 - m^2) f_\beta(y) \quad (8.12)$$

By standard manipulations, ∇^2 on f_β is transferred, by partial integrations and dropping contributions at spatial infinity, to ∇^2 acting on the matrix element. While this is a standard procedure, the reader should, nevertheless, check that it is justified.

Putting everything together, one gets a very compact expression for the non-trivial part of $\mathcal{M}$:

$$i \int dy f_\beta(y) \mathcal{K}_y \langle \alpha' \beta', out | \phi(y) | \alpha \rangle \quad (8.13)$$

Here, $\mathcal{K}_y$ is the Klein-Gordon operator ($\square + m^2$), and it is important to note that it acts only on the matrix element, and not on $f_\beta(y)$. It should be noted that this expression is completely covariant even though we used non-covariant methods in its derivation. The idea is to repeatedly apply this “reduction” to as many particles as one wishes. To avoid clutter, let us introduce

$$\mathcal{N}(y) \equiv \langle \alpha' \beta', out | \phi(y) | \alpha \rangle \quad (8.14)$$

so that $\mathcal{M}$ (the non-trivial part of it) reads

$$\mathcal{M} = i \int dy \{ \mathcal{K}_y \mathcal{N}(y) \} f_\beta(y) \quad (8.15)$$

We shall adopt the convention that the f ’s for the incoming particles are to the right, and those of the outgoing to the left. At this stage, there are three states one could reduce: α' , β' , α . Following Barton, we shall reduce the outgoing β' state, though, as emphasized, there is complete freedom in this regard.

Consequently,

$$\begin{aligned} \mathcal{N}(y) &= \langle \alpha' | a_{\beta'}^{(out)} \phi(y) | \alpha \rangle \\ &= i \int d\mathbf{x} f_{\beta'}^*(x) \overleftrightarrow{\frac{\partial}{\partial x_0}} \langle \alpha' | \phi(x) \phi(y) | \alpha \rangle |_{x_0 \rightarrow \infty} \end{aligned} \quad (8.16)$$

We have used the weak asymptotic condition. However, strictly speaking, there is a potentially troublesome technicality at this point. The weak asymptotic condition was formulated for matrix elements of the type $\langle A | \phi(x) | B \rangle$, where $|A\rangle$, $|B\rangle$ are arbitrary particle states (not necessarily single particle). But in the previous step, we have used it for $\langle \alpha' | \phi(x) \phi(y) | \alpha \rangle$. In principle, depending on the circumstances, $\phi(y) | \alpha \rangle$ could involve an *infinite* superposition of particle states and the interchange of the $x_0 \rightarrow \infty$ and the summation may not be legitimate.

At this point, almost every source indulges in some tricks which are claimed to be *harmless*. Let us take up the first, and the most frequently used. This consists in the replacement

$$\begin{aligned} \phi(x) \phi(y) &\xrightarrow{x_0 \rightarrow \infty} T(\phi(x) \phi(y)) \\ &\equiv \theta(x_0 - y_0) \phi(x) \phi(y) + \theta(y_0 - x_0) \phi(y) \phi(x) \end{aligned} \quad (8.17)$$

The replacement certainly seems legitimate. Making use of

$$\int_{x_0 \rightarrow -\infty} d\mathbf{x} = \int_{x_0 \rightarrow \infty} d\mathbf{x} - \int dx \frac{\partial}{\partial x_0} \quad (8.18)$$

We split the contributions into two parts, (i) the $x_0 \rightarrow -\infty$, which shall be called the *boundary part* and (ii) the rest, involving the four-dimensional integral over $\mathbf{x}$, which we shall call the *bulk part*. The boundary contribution to $\mathcal{N}$ is

$$\mathcal{N}_{bnd}(y) = \int d\mathbf{x} f_{\beta'}^*(x) \frac{\overleftrightarrow{\partial}}{\partial x_0} \langle \alpha' | \phi(y) \phi(x) | \alpha \rangle |_{x_0 \rightarrow -\infty} \quad (8.19)$$

where we have already evaluated the T-product at $x_0 \rightarrow -\infty$. On using the asymptotic condition on $\phi(x)$,

$$\mathcal{N}_{bnd}(y) = i \int d\mathbf{x} f_{\beta'}^*(x) \frac{\overleftrightarrow{\partial}}{\partial x_0} \langle \alpha' | \phi(y) \phi_{in}(x) | \alpha \rangle \quad (8.20)$$

Using the expression for $a_{\beta', in}$, this becomes

$$\mathcal{N}_{bnd}(y) = \langle \alpha' | \phi(y) a_{\beta', in} | \alpha \rangle = \delta_{\alpha\beta'} \langle \alpha' | \phi(y) | 0 \rangle = \delta_{\alpha\beta'} f_{\alpha'}(y) \quad (8.21)$$

Our interest is in $\mathcal{M}$ and as per Eq.(8.15), this boundary contribution vanishes because $\mathcal{K}_y f_{\alpha'}(y) = 0$. Thus, the replacement by T-product has the dual advantages of agreement at the $x_0 \rightarrow \infty$, as well as of no boundary contributions at $x_0 \rightarrow -\infty$. Let us complete the calculation of the bulk contribution to $\mathcal{N}$ first, and finally to $\mathcal{M}$:

$$\mathcal{N}_{bulk}(y) = i \int dx \frac{\partial}{\partial x_0} \{ f_{\beta'}^*(x) \frac{\overleftrightarrow{\partial}}{\partial x_0} \langle \alpha' | T(\phi(x) \phi(y)) | \alpha \rangle \} \quad (8.22)$$

By using the same kind of manipulations used earlier for the y -variable, it is easy to see that this becomes

$$\mathcal{N}_{bulk}(y) = i \int dx f_{\beta'}^*(x) \{ \mathcal{K}_x \langle \alpha' | T(\phi(x) \phi(y)) | \alpha \rangle \} \quad (8.23)$$

and, for $\mathcal{M}$ (henceforth, we shall always the non-trivial part of $\mathcal{M}$):

$$\mathcal{M} = (i)^2 \int \int dx dy f_{\beta'}^*(x) \{ \mathcal{K}_x \mathcal{K}_y \langle \alpha' | T(\phi(x) \phi(y)) | \alpha \rangle \} f_{\beta}(y) \quad (8.24)$$

It is easy to recognize a pattern in all this, and to write down the answer when all four particles are reduced:

$$\mathcal{M} = (i)^4 \int \prod_{i=1}^4 dx_i f_{\alpha'}^*(x_1) f_{\beta'}^*(x_2) \{ \prod_{i=1}^4 \mathcal{K}_i \langle 0 | T(\prod_{j=1}^4 \phi(x_j) | 0) \rangle \} f_{\alpha}(x_3) f_{\beta}(x_4) \quad (8.25)$$

Here, $T(\phi(x_1)\phi(x_2)\phi(x_3)\phi(x_4))$ is an obvious generalization of the T-product for two fields introduced earlier. Generalizations of the above to cases where there are arbitrary number of incoming particles to arbitrary number of outgoing particles is also straightforward. There are a number of important remarks that we wish to make about this representation:

- This is a fully non-perturbative result unlike, for example, Dyson's S-matrix discussed earlier which was entirely perturbative.
- This, and generalizations to more incoming and outgoing particles, shows that the entire observable content of any RQFT is coded in the vacuum expectation values of the T-products of fields. We shall shortly present another equivalent representation in terms of so-called *Retarded Commutators*.
- This non-perturbative derivation has made no use of any explicit Lagrangeans, Hamiltonians or ETCR's. Though superficially it looks very different from Dyson's S-matrix in QED, it is completely equivalent to it. The Dyson operator explicitly depended on the interaction Hamiltonian. In the LSZ form, interaction dependence enters via the action of $\mathcal{K}$'s on the field operators. In the extreme limit of no interactions, the field operators are annihilated by these leading to no non-trivial scatterings.
- A very important property of the LSZ form is that all dependences on the external momenta are through the f_α 's. This allows for a direct analysis of the analytical properties of S-matrix elements, as will be discussed in detail in the next chapter.
- The generalizations of these results to the physically more relevant cases of fermions and bosons of unequal masses are also straightforward and the reader is referred to the books of Gasiorowicz [3] and Bjorken-Drell [2].

8.1.1 The Retarded Commutator Representation

The previous discussion raises the interesting question of whether there are other replacements or tricks that have all the advantages of the T-product replacement. The answer is indeed in the affirmative. To see this let us begin with the initial expression for $\mathcal{N}(y)$:

$$\mathcal{N}(y) = \langle \alpha' \beta', out | \phi(y) | \alpha \rangle = \langle \alpha' | a_{\beta', out} \phi(y) | \alpha \rangle \quad (8.26)$$

Before proceeding along the lines done previously, we rewrite this expression as

$$\mathcal{N}(y) = \langle \alpha' | [a_{\beta', out}, \phi(y)] | \alpha \rangle + \langle \alpha' | \phi(y) a_{\beta', out} | \alpha \rangle \quad (8.27)$$

The second term does not contribute to $\mathcal{M}$ for reasons we have seen before: treating $|\alpha\rangle$ as $|\alpha, out\rangle$ by invoking single particle stability, this term becomes $\delta_{\alpha\beta'} \langle \alpha' | \phi(y) | 0 \rangle = \delta_{\alpha\beta'} f_{\alpha'}^*(y)$, and $\mathcal{K}_y f_{\alpha'}^*(y) = 0$. This means, a new representation for $\mathcal{N}$, which we shall call *Retarded Commutator Representation*, can also be considered:

$$\mathcal{N}_{ret}(y) = \langle \alpha' | \theta(x_0 - y_0) [\phi(x), \phi(y)] | \alpha \rangle |_{x_0 \rightarrow \infty} \quad (8.28)$$

It is immediately clear that the boundary contribution to $\mathcal{N}_{ret}(y)$ at $x_0 \rightarrow -\infty$ vanishes identically as the $\theta(x_0 - y_0)$ vanishes there. Thus the retarded commutation representation has the same dual advantages as the T-product representation. We shall

simply present the final result for the bulk contribution, as the manipulations leading to it should be fully familiar by now.

$$\mathcal{M}_{ret} = (i)^2 \int \int dx dy f_{\beta'}^*(x) \{ \mathcal{K}_x \mathcal{K}_y \langle \alpha' | [\phi(x), \phi(y)]_{ret} | \alpha \rangle \} f_{\beta}(y) \quad (8.29)$$

The retarded product can easily be generalized to the case of n fields, but we shall not present it here.

The equality of the two representations: it is clear that at least heuristically, the T-product and the R-product representations should yield the same scattering amplitudes. But as Barton has commented, their operator forms are so different that a direct check of this equality may be worthwhile. On noting that

$$\begin{aligned} T(\phi(x)\phi(y)) - R(\phi(x)\phi(y)) &= \theta(x_0 - y_0)\phi(x)\phi(y) + \theta(y_0 - x_0)\phi(y)\phi(x) \\ &\quad - \theta(x_0 - y_0)[\phi(x), \phi(y)] = \phi(y)\phi(x) \end{aligned} \quad (8.30)$$

Since no θ -function occurs in the relevant product, the Klein-Gordon operators $\mathcal{K}_x, \mathcal{K}_y$ can directly act on the fields $\phi(x), \phi(y)$ to give the *sources* $\eta(x) \equiv \mathcal{K}_x \phi(x)$ and $\eta(y) \equiv \mathcal{K}_y \phi(y)$. Consequently

$$\delta = \mathcal{M}_T - \mathcal{M}_R = \frac{(i)^2}{\sqrt{2\omega_{b'}} 2\omega_b} \int \int dx dy e^{i k_{b'} x - i k_b y} \langle \alpha' | \eta(y) \eta(x) | \alpha \rangle \quad (8.31)$$

On introducing a complete set of states and using $\phi(x) = e^{iPx} \phi(0) e^{-iPx}$ with P the momentum operator, and carrying out the x, y integrations, it is easy to see that the above is

$$\delta = \frac{(i)^2}{\sqrt{2\omega_{b'}} 2\omega_b} \sum_n \langle \alpha' | \eta(0) | n \rangle \langle n | \eta(0) | \alpha \rangle \delta(k_{b'} - k_a + k_n) \delta(k_{a'} - k_b - k_n) \quad (8.32)$$

Single-particle stability leads to a vanishing δ . There are two parts to this argument, and Barton discusses only the necessary, but not sufficient condition. Take either of the delta functions (overall energy-momentum conservation implies that one of them implies the other), say, $k_a = k_{b'} + k_n$. But this is the energy-momentum conservation condition for the reaction $a \rightarrow b + n$ and single particle stability will not allow it unless $k_n = 0$, i.e. “ n ” is the vacuum state. Single particle stability, stated in terms of the sources is $\langle 0 | \eta(0) | k \rangle = 0$, and this has the consequence that $\delta = 0$. But it is clear that even if the energy-momentum conservation is satisfied, the process may not be allowed because no source connects the initial “ a ” state with the final “ $b + n$ ”. That would also imply $\langle \alpha' | \eta(0) | n \rangle = 0$, which again leads to $\delta = 0$. Therefore the T-product and R-product representations indeed lead to the same scatterings.

8.1.2 What if No Tricks are Used?

Even though the replacement of $\phi(x)\phi(y)$ as $x_0 \rightarrow \infty$ by either $T(\phi(x)\phi(y))$ or $\theta(x_0 - y_0) [\phi(x), \phi(y)]$ has been argued to be both legitimate and desirable, it still does not satisfactorily answer the question “why perform the trick at all, why not leave the original product as it is?”. We answer this by a careful analysis. The bulk part of $\mathcal{M}$ would then be:

$$\mathcal{M}_{bulk}^{notrick} = (i)^2 \int \int dx dy f_{\beta'}^*(x) \{ \mathcal{K}_x \mathcal{K}_y \langle \alpha' | \phi(x)\phi(y) | \alpha \rangle \} f_{\beta}(y) \quad (8.33)$$

If anything, this seems to have the advantage that the Klein-Gordon operators can directly operate on the fields to yield the sources, without the θ -functions coming in the way. Likewise, the boundary part is given by

$$\mathcal{M}_{bnd}^{notrick} = - \int dy d\mathbf{x} f_{\beta'}^*(x) \{ \mathcal{K}_y \frac{\overleftrightarrow{\partial}}{\partial x_0} \langle \alpha' | \phi(x)\phi(y) | \alpha \rangle \} f_{\beta}(y)|_{x_0 \rightarrow -\infty} \quad (8.34)$$

There is clearly no reason why this should vanish in general. This answers what happens if no tricks were adopted. In general, the boundary terms do not vanish, and what is more, the boundary terms are not manifestly covariant. It should be emphasized that there is nothing wrong in principle with this. It is just that the resulting expressions are not particularly useful for further applications like the covariant calculus for S-matrix elements or the non-perturbative analysis of the analyticity properties of scattering amplitudes.

8.1.3 Unretarded Commutator Representation

At $x_0 \rightarrow \infty$, $\phi(x)\phi(y)$ could equally well have been replaced by $\theta(x_0 - y_0) [\phi(x), \phi(y)]$ or by $[\phi(x), \phi(y)]$. The latter is called the *Unretarded Commutator*. We now explore the consequences of this unretarded commutator representation. At $x_0 \rightarrow \infty$, it produces the same contribution as if no tricks had been performed, i.e. as if one had simply continued with $\phi(x)\phi(y)$ in place of either the T -product or the retarded commutator. That leaves the question of the boundary contribution. Now, there is an additional contribution of a $-\phi(y)\phi(x)$ to the boundary over and above that of the untampered $\phi(x)\phi(y)$ term. But this is identical (apart from a sign) to the boundary contribution encountered in the T -product case that was shown not to contribute. Therefore, the unretarded commutator representation and the untampered case give the same scattering amplitudes. Both of them suffer from nonvanishing boundary terms.

Though the unretarded representation is not as attractive as either the T -product or the retarded commutator, we shall see in the next chapter that it provides a crucial link in non-perturbatively establishing the analyticity properties of scattering amplitudes. The reason for this is that the analyticity properties of *unretarded commutator* representation is straightforward to establish. Though the LSZ reduction

results in a *retarded* representation, there is a crucial link between the momentum representations of the retarded and unretarded commutators.

8.1.4 Crossing Symmetry

One of the spectacular spin-offs from the LSZ approach is a non-perturbative derivation of the so-called *Crossing Relations*. This is a crucial ingredient to the analytic S-matrix and Dispersion Relations program to strong interaction physics, but it nevertheless has to be postulated ad hoc. Only RQFT yields the crossing relations as a consequence. Though these can be formulated even within the single neutral scalar field theory we have used so far to illustrate the LSZ reductions, we shall consider two simple generalizations to bring out the full power of the crossing relations. The extensions we consider are (i) two distinct neutral scalar fields and (ii) an extension where *particles and their antiparticles are distinct*.

Right at the outset it helps to set the notion of *Physical Regions* for scattering. Among other things, the energies of both the ingoing and outgoing particles have to be *positive* in order for the scattering to be a physical process. For the moment, this suffices. In the next chapter, we shall bring out other restrictions that physical regions must satisfy. Now let us introduce two distinct particles A and B (and all their associated fields) with masses m_a, m_b . Let us first consider a *physical* scattering process:

$$A(k_a) + B(k_b) \rightarrow A(k'_a) + B(k'_b) \quad (8.35)$$

This means $(k_{a,0}, k_{b,0}, k'_{a,0}, k'_{b,0})$ are *all* positive. Now, let us consider a totally different class of scatterings, also physical:

$$A(k_a) + A(k''_a) \rightarrow B(k'_b) + B(k''_b) \quad (8.36)$$

For convenience, we have chosen the momenta k_a, k'_b to be common to both the processes. This is really not necessary. Again, for the newly introduced momenta, $k_{a'',0}, k_{b'',0}$ are both positive. It is quite obvious that these two classes of scattering processes are quite distinct, with no obvious connection. Rather remarkably, the LSZ formalism relates these two processes, albeit in a nuanced manner. The nuance being it relates the scattering amplitude for a *physical* process of one kind to an *unphysical extrapolation* of the other.

For explicit details of this connection based on the LSZ reductions of the type we obtained earlier, the reader is referred to Barton's book. What the LSZ formalism does is to relate the physical process of Eq. (8.35) to the unphysical process

$$A(k_a) + A(-k_{a'}) \rightarrow B(k_{b'}) + B(-k_b) \quad (8.37)$$

Though this too satisfies the same overall energy-momentum conservation $k_a - k_{a'} = k_{b'} - k_b$ as the physical process of Eq. (8.35) i.e $k_a + k_b = k_{a'} + k_{b'}$, this process can not be physically realized because $-k_{a',0} < 0$ and so is $-k_{b,0}$. What we

meant by the LSZ formalism relating these processes can be given a very precise meaning: consider the physical process of Eq. (8.36), whose energy-momentum conservation reads $k_a + k_{a''} = k_b + k_{b''}$, and in the expression for the LSZ scattering amplitude, make the substitutions $k_{a''} \rightarrow -k_{a'}$, $k_{b''} \rightarrow -k_b$. Note that the energy-momentum conservation with these substitutions is still a valid condition, coinciding with the energy-momentum conservation law for the *physical* process of Eq. (8.35). Now the LSZ formalism says that this latter amplitude, extrapolated to an unphysical region for scatterings of the type $A + A \rightarrow B + B$, is exactly *equal* to the physical scattering amplitude for process $A + B \rightarrow A + B$.

Since the relation is between one process in its physical region, and another in its unphysical region, it can not be tested directly experimentally. But in the analytic S-matrix and Dispersion relations approach to strong interactions (a lot more on this as we go along), one has analytic continuations to unphysical regions and crossing symmetry relations become powerful constraints on the allowed analytic continuations.

Now we discuss the crossing symmetric relations in a broader context, in theories where particles and their *antiparticles* are distinct. This can happen, for example, in complex scalar field theories, as also in Fermionic field theories. Details of spin, etc., are not important for this discussion. Then let us consider processes of the type

$$A(p_a) + B(p_b) \rightarrow C(p_c) + D(p_d) \quad (8.38)$$

which is physical, i.e. $(p_{a,0}, p_{b,0}, p_{c,0}, p_{d,0})$ are all positive. Denoting the antiparticles of C,D by $\bar{C}, \bar{D}$, let us consider the process

$$A(p_a) + \bar{C}(q_{\bar{c}}) \rightarrow \bar{B}(q_{\bar{b}}) + D(p_d) \quad (8.39)$$

also in its physical region, i.e. when $q_{\bar{c},0}, q_{\bar{b},0}$ are both positive. Like before, this process in the *unphysical* region

$$A(p_a) + \bar{C}(-p_c) \rightarrow \bar{B}(-p_b) + D(p_d) \quad (8.40)$$

is related to the physical process of Eq. (8.38). Thus the LSZ formalism is so powerful that it not only yields crossing relations between scatterings of only particles but also yields crossing relations between scatterings involving both particles and antiparticles. In other words, it relates particle creation, particle annihilation and scatterings among particles and antiparticles. The essential reason for this unity is that all these are welded into the same *Quantum Field*.

This is a good point to revisit our discussion of the Kramers-Kronig relations. There, a so-called symmetry relation relating the dielectric constant at *negative frequency* to the complex conjugate at positive frequency was essential to a derivation of the Dispersion Relations. The reader is referred to Eq. (4.10). This is the exact analog of the crossing relations discussed here, with negative frequencies playing the role of the unphysical regions.

References

1. G. Barton, *Introduction to Advanced Field Theory*, *Interscience Tracts on Physics and Astronomy*, 22 (1963)
2. J.D. Bjorken, S. Drell, *Relativistic Quantum Fields* (McGraw Hill Publishers)
3. S. Gasiorowicz, *Elementary Particle Physics* (Wiley)

9.1 Introduction

In this chapter, we address the issue of unitarity as a prelude to analyticity of scattering amplitudes as a consequence of RQFT. There are many aspects of such analytic behaviour. One such is to derive the analogs of the Kramers-Kronig relations of optics, but now in RQFT. In optics, this arose from the analyticity properties of the *Dielectric Function* as a function of *Complex Frequencies*. In field theories, this becomes the question of analytic behaviour of scattering amplitudes, equivalently, S-matrix elements, when one or more of their arguments are continued away from their physical values. The continuation can be to both the cases of these variables assuming real but unphysical values, or to complex values. In the optics case, the frequency taking real but negative values is continuation of the first type, while frequencies continued to complex values are of the latter kind.

The situation with scattering amplitudes is much more complicated. Even for the simplest case of two initial particles scattering into two final ones, there are two invariants on which the amplitudes depend on (more details shortly), and it becomes a problem of the analytic behaviour of a function of *two* complex variables, which is considerably harder (euphemistically called “much richer”!) than the corresponding problem of functions of a single complex variable. In the optics case, the physical region was simply that of frequencies being positive, i.e. $\omega \geq 0$ (zero frequency corresponds to the static case). Even in the simplest $2 \rightarrow 2$ scattering, there are three invariants (s, t, u) subject to the constraint $s + t + u = \sum_{i=1}^4 m_i^2$. There are *three* physical channels, the so-called s, t, u -channels each with its own physical region. Any of the two independent variables can be chosen to be the “coordinates” of the kinematical space. If we choose (s, t) , then there are three *physical regions* for the three channels. Everything outside them constitutes the overall unphysical region though, for each channel, even the physical regions of the other two channels are unphysical. The ultimate challenge is to find the maximally analytic function of

the two complex variables (s, t) such that its boundary values for the three physical regions gives the physical scattering amplitudes for the corresponding channels. By maximally analytic, one means analytic everywhere except for poles and cuts as demanded by physical requirements of unitarity, crossing symmetry, etc.

The main goals of such Kramers-Kronig-type analyticity investigations are the so-called *Dispersion Relations* relating the real part of the scattering amplitudes to their imaginary parts, i.e. Hilbert Transforms. The ones that have predictive powers are the cases where the imaginary or *absorptive parts* are directly measurable. This is the situation with the so-called *Forward Scattering Dispersion Relations*. Such forward scattering dispersion relations are straightforward to prove when one of the particles is *massless* as, for example, in the scattering of photons. For massive particles, however, even the forward scattering dispersion relations are tricky to prove. Symanzik [1], as well as Bogoliubov, Jost and Lehmann (as cited by Symanzik in [1]) were among the first to prove forward scattering dispersion relations. Away from the forward scattering, proofs of such dispersion relations become increasingly complicated. Capps and Takeda [2], and Salam [3] were among the first to prove them. Chew, Goldberger, Low and Nambu [4,5] applied such fixed- t dispersion relations to the study of pion-nucleon scattering as well as to photoproduction of pions. The paper by Capps and Takeda has references to many important works.

Consequences of analyticity of scattering amplitudes are many, and very important. Some of the most notable ones are (i) the Froissart bound for the high-energy behaviour of scattering amplitudes, (ii) the Jin-Martin bound, (iii) the equality of π^+p and π^-p total cross sections at very high energies, absolute bounds on $\pi - \pi$ -scattering amplitudes, etc. The reader is referred to a short but very succinct account by Andre Martin [6], and the lectures by Khuri in the First Brookhaven Summer School [7]. These are also covered in Chap. 12 of this book.

In this chapter, we shall however focus on analyticity properties of scattering amplitudes as a function of *momentum transfer*. The reason is that, on the one hand, these consequences follow immediately from the LSZ reduction formulae obtained in the previous chapter, and the derivations by Lehmann [8] are so transparent that even beginning students will readily appreciate how RQFT leads to powerful statements on analyticity and unitarity. On the other hand, the forms of analyticity so obtained, the so-called Lehmann-Ellipses, provide a smooth transition to what interests us majorly in this book like analytic continuation of *Partial wave expansions*, Regge Poles, etc. We shall return to dispersion relations in Chap. 11.

Analyticity in momentum transfer inside the Lehmann ellipses can, with some additional consideration of analyticity in the *masses* of the external particles, lead to direct proofs of fixed- t dispersion relations. We shall return to this later when we take up dispersion relations. A very readable account of this can be found in Lehmann's lecture notes *Scattering Matrix and Field Operators* [9]. A detailed, and

also somewhat terse, account can be found in Sommer's review article *Present State of Rigorous Analytic Properties of Scattering Amplitudes* [10].¹

9.2 General Considerations

Particle masses play greater role in the discussion of analyticity than, say, spin, charges and other internal quantum numbers (they are important too, at the level of details). Lehmann considers two species of particles (fields) with unequal masses. Though he calls them *nucleons* and *mesons*, they are both taken to be *spin-less*. Apart from introducing an additional field, all the techniques from the previous chapter can be taken over. Before using scattering amplitudes so obtained, a simple but very important modification to the form of these scattering amplitudes is in order. It is easier to demonstrate this for the case of just one field, as was considered in the last chapter, as generalizations to the case of more fields are straightforward.

The Mandelstam Variables This is a good place to introduce the *Mandelstam Variables*. For the general scattering process,

$$A(p_A) + B(p_B) \rightarrow C(p_C) + D(p_D) \quad p_A + p_B = p_C + p_D \quad (9.1)$$

The Mandelstam variables are

$$s = (p_A + p_B)^2 \quad t = (p_A - p_C)^2 \quad u = (p_A - p_D)^2 \quad (9.2)$$

For the most general case where $p_A^2 = m_A^2$, $p_B^2 = m_B^2$, $p_C^2 = m_C^2$, $p_D^2 = m_D^2$ with all the four different masses, it is an elementary exercise to show

$$s + t + u = m_A^2 + m_B^2 + m_C^2 + m_D^2 \quad (9.3)$$

We saw that there are many different forms for the scattering amplitudes in the LSZ formalism. These differences arise, on the one hand, from which of the particle states are *reduced*, and on the other hand from the use of T-products of fields or of the *retarded product* representations. For the purposes of determining analyticity properties, the retarded product representations have so far proved to be more useful. Let us start by recalling the retarded product representation when one particle from the initial state and one particle from the final state have been reduced as in Eq. (8.29) (the contribution of the *trivial part* of the S-matrix, i.e. the identity, has been left out) (the k 's are the momenta of the β particles):

$$\begin{aligned} \mathcal{M}_{ret} &\equiv \langle \alpha' \beta', out | \alpha \beta, in \rangle \\ &= (i)^2 \int \int dx dy e^{ik_{b'} \cdot x} \{ \mathcal{K}_x \mathcal{K}_y \langle \alpha' | [\phi(x), \phi(y)]_{ret} | \alpha \rangle \} e^{-i k_b \cdot y} \end{aligned} \quad (9.4)$$

¹ I am indebted to Jnanadeva Maharana for an extensive discussion on this, and for bringing my attention to these proofs.

We have taken the plane-wave limit for the f s. To avoid clutter, let us introduce the shorthand

$$\mathcal{O}(x, y) = \mathcal{K}_x \mathcal{K}_y \theta(x_0 - y_0) [\phi(x), \phi(y)] \equiv i R' \{\phi(x) \phi(y)\} \quad (9.5)$$

where we have introduced the *Retarded Commutator* R' of Lehmann [8]. Because of the presence of the θ -function, and that the Klein-Gordon operators $\mathcal{K}_x, \mathcal{K}_y$ act on both the θ -function, as well as the field products, the retarded commutators R' so introduced are not even commutators. In his lectures [9], Lehmann introduces another object, the R-product:

$$R\{\phi(x)\phi(y)\} \equiv -i \theta(x_0 - y_0) [\phi(x), \phi(y)] \quad (9.6)$$

which is indeed the retarded commutator. Despite the somewhat misleading notation, R and R' are very different objects; they do not even have the same mass dimensions! Next, one introduces the *T-matrix* (we introduce the shorthand notation $\delta^{(4)}(k_{tot})$ for the overall energy-momentum conservation):

$$\langle \alpha' \beta', out | \alpha \beta, in \rangle \equiv \langle \alpha' \beta', in | \alpha \beta, in \rangle + i (2\pi)^4 \delta^{(4)}(k_{tot}) T(\alpha \beta, \alpha' \beta') \quad (9.7)$$

It is worth pointing out that it is the T -matrix which is more useful in all manipulations with scattering amplitudes. The representation used by Lehmann for the T -matrix (which we denote by T_L) in proving various analyticity properties, is, for example,

$$T_L(\alpha \beta; \alpha' \beta') \equiv - \int dx e^{i \frac{k_{\beta'} + k_{\beta}}{2} \cdot x} \left\langle \alpha' | R' \left\{ \phi\left(\frac{x}{2}\right) \phi\left(-\frac{x}{2}\right) \right\} | \alpha \right\rangle \quad (9.8)$$

Despite their totally different mathematical forms, we shall show that T_L of Eq. (9.8) as defined by Lehmann, and T as obtained from the LSZ formalism according to Eqs. (9.7) and (9.8) are indeed the same, i.e. $T_L = T$. Let us take a closer look at their obvious difference in forms. Equation (9.4) is a *double* space-time integral over x, y . Equation (9.8), on the other hand, is a *single* space-time integral over x . One part of the S-matrix element is in a sense trivial, but very important. That can be deduced without any detailed calculations, and that is the delta-function expressing the overall conservation of energy-momentum. The T-matrix on the other hand is what obtains after this delta-function has already been factored out. This makes it evident that the role of an extra space-time integration in Eq. (9.4) is just to produce the overall energy-momentum conservation delta-function. Instead of figuring out the requisite linear transformations among x, y that would result in a clear demonstration of this, we shall follow a different strategy.

Before that let's take a look at the other major difference. In Eq. (9.4) what occurs is $R' \{\phi(x) \phi(y)\}$ which is a function of *two* space-time coordinates x, y ; on the other hand, Eq. (9.8) involves $R' \{\phi(\frac{x}{2}) \phi(-\frac{x}{2})\}$ which involves only one space-time vector x . How is this latter quantity to be interpreted? If one naively substitutes $x \rightarrow \frac{x}{2}$ and $y \rightarrow -\frac{x}{2}$, then one would end up $\mathcal{K}_{\frac{x}{2}}^2$ which is obviously incorrect. The correct

interpretation is to first determine $R'\{\phi(x)\phi(y)\}$ as a function of two variables x, y , and then evaluate this function of two variables at the special point $(\frac{x}{2}, -\frac{x}{2})$. More precisely,

$$R'\{\phi(\frac{x}{2})\phi(-\frac{x}{2})\} = \int \int dx' dy' \delta^{(4)}(x' - \frac{x}{2}) \delta^{(4)}(y' + \frac{x}{2}) R'\{\phi(x')\phi(y')\} \quad (9.9)$$

It is important to appreciate the subtlety behind this distinction. On using this, we can recast T_L as

$$T_L = - \int \dots \int dx dx' dy' e^{i \frac{(k_\beta + k_{\beta'})}{2} \cdot x} \delta^{(4)}(x' - \frac{x}{2}) \delta^{(4)}(y' + \frac{x}{2}) \langle \alpha' | R'\{\phi(x)\phi(y)\} | \alpha \rangle \quad (9.10)$$

One can view Eq. (9.4) as a double Fourier-transform of $i R'\{\phi(x)\phi(y)\}$. On inverse Fourier transforming (we momentarily introduce the shorthand notation $\bar{T}(\{k_i\}; \{k_f\}) = \delta^{(4)}(\sum k_i - \sum k_f) T(\{k_i\}; \{k_f\})$ to manage the length of the equations)

$$\langle \alpha' | R'\{\phi(x)\phi(y)\} | \alpha \rangle = - \frac{1}{(2\pi)^4} \int \int dk_\gamma dk_{\gamma'} e^{ik_{\gamma'} \cdot y' - ik_{\gamma'} \cdot x'} \bar{T}(k_\alpha k_\gamma; k_{\alpha'} k_{\gamma'}) \quad (9.11)$$

To bring greater clarity, let us evaluate T_L in steps. On substituting the double inverse FT in the previous expression,

$$T_L = \frac{1}{(2\pi)^4} \int \int dx dx' dy' dk_\gamma dk_{\gamma'} e^{i \frac{(k_\beta + k_{\beta'})}{2} \cdot x - i k_{\gamma'} \cdot x' + i k_{\gamma'} \cdot y'} \delta^{(4)}(x' - \frac{x}{2}) \cdot \delta^{(4)}(y' + \frac{x}{2}) \delta^{(4)}(k_\alpha + k_\gamma - k_{\alpha'} - k_{\gamma'}) T(k_\alpha k_\gamma; k_{\alpha'} k_{\gamma'}) \quad (9.12)$$

Performing the x', y' integrations

$$T_L = \frac{1}{(2\pi)^4} \int \int \int dx dk_\gamma dk_{\gamma'} e^{i(\frac{k_\beta + k_{\beta'}}{2} - \frac{k_{\gamma'} + k_{\gamma'}}{2}) \cdot x} \bar{T}(k_\alpha k_\gamma; k_{\alpha'} k_{\gamma'}) \quad (9.13)$$

Finally, the x integration can be done to yield

$$T_L = \int \int dk_\gamma dk_{\gamma'} \delta^{(4)}(k_\beta + k_{\beta'} - k_\gamma - k_{\gamma'}) \delta^{(4)}(k_\alpha - k_{\alpha'} + k_\gamma - k_{\gamma'}) \cdot T(k_\alpha k_\gamma; k_{\alpha'} k_{\gamma'}) \quad (9.14)$$

The two delta functions yield $k_\beta + k_{\beta'} = k_\gamma + k_{\gamma'}$, and, $k_\gamma - k_{\gamma'} = k_{\alpha'} - k_\alpha$. The solution to these constraints is: $k_\gamma = k_\beta$ and $k_{\gamma'} = k_{\beta'}$. Consequently,

$$T_L = T \quad (9.15)$$

Similar equalities can be obtained for other T -matrix elements.

9.3 LSZ Formalism and Unitarity

The scattering matrix, S , which can also be viewed as the matrix transforming arbitrary out-states into in-states, must be *Unitary*. We set the notation by explicitly showing the transformation

$$\langle \alpha', out | \alpha, in \rangle \equiv \langle \alpha', in | S | \alpha, in \rangle \rightarrow S^\dagger | \alpha', in \rangle = | \alpha', out \rangle \quad (9.16)$$

(see, for example, Eq. (4.1) of [11], or, Eq. (6.8) of [12]).

We saw in the last chapter how the LSZ formalism automatically accounts for *Crossing Symmetry*. Now we shall show that the scattering amplitudes given by the LSZ formalism also satisfy *unitarity* automatically. This speaks for the power and economy of the LSZ formalism! Though unitarity can be proved for arbitrary scattering amplitudes obtained from the LSZ formalism, we shall explicitly demonstrate it for the case of $2 \rightarrow 2$ scattering of a single species of spin-less particles. Further, it is easier to show this in the version with double space-time integrals. The reader should have no difficulty in making appropriate generalizations.

Before getting into the details, let us state the unitarity conditions in terms of the relevant T -matrices. At the *Operator* level,

$$\mathbf{S} = \mathbf{1} - i\mathbf{T} \quad (9.17)$$

Unitarity of S -matrix requires $\mathbf{S}^\dagger \mathbf{S} = \mathbf{1} = \mathbf{S} \mathbf{S}^\dagger$. It suffices to consider any of these two, so let us analyse the first.

$$\mathbf{S}^\dagger \mathbf{S} = \mathbf{1} \rightarrow (\mathbf{1} + i\mathbf{T}^\dagger)(\mathbf{1} - i\mathbf{T}) = \mathbf{1} \rightarrow i(\mathbf{T}^\dagger - \mathbf{T}) = -\mathbf{T}^\dagger \mathbf{T} \quad (9.18)$$

This has to be rewritten in terms of matrix elements. The matrix elements of the T -operator have the general structure

$$\mathbf{T}_{ij} = \langle i | \mathbf{T} | j \rangle = (2\pi)^4 \delta^{(4)}(P_i - P_j) T(i; j) \quad (9.19)$$

Thus, in $T(i; j)$, j is the initial state and i is the final state. With these preliminaries, Eq. (9.18) can be written in terms of the matrix elements $T(i; j)$ as

$$\begin{aligned} & i(2\pi)^4 \delta^{(4)}(p + q - p' - q')(T^*(pq; p' q') - T(p' q'; pq)) \\ &= - \sum_n \langle p' q' | \mathbf{T}^\dagger | n \rangle \langle n | \mathbf{T} | pq \rangle \\ &= -(2\pi)^8 \delta^{(4)}(p' + q' - p - q) \sum_n \delta^{(4)}(p_n - p - q) T^*(n; p' q') T(n; pq) \end{aligned} \quad (9.20)$$

In arriving at this, a complete set of intermediate states is introduced. There is one energy-momentum conservation delta function from each of the matrix elements, and, not surprisingly, the two delta functions can be written as a product of the

overall delta function, another which forces the p_n of the intermediate state to be $p + q(p' + q')$. The above can be simplified to

$$i(T^*(pq; p'q') - T(p'q'; pq)) = -(2\pi)^4 \sum_n \delta^{(4)}(p_n - p - q) T^*(n; p'q') T(n; pq) \quad (9.21)$$

This is expressing the unitarity constraint in terms of the T -matrix elements; it is not a proof of the unitarity relation, which we now proceed to prove within the LSZ formalism. A few clarificatory remarks are in order at this stage. The $T(i; j)$ are generally *complex*, and talking about their imaginary part $\text{Im } T(i; j)$ certainly makes sense. At an operator level, $\mathbf{T}^\dagger - \mathbf{T}$ can be thought of as $2i \text{Im} \mathbf{T}$. Sometimes, the matrix element of this imaginary part of the operator is mixed up with the imaginary part of a T -matrix element. One should make a careful distinction between the two. Only in the case of *forward scattering* are they the same. In fact, the matrix element of $\text{Im} \mathbf{T}$ necessarily involves two distinct T -matrix elements, as can be clearly seen from Eq. (9.21). However, it is the latter that will be seen to play the important role of the *absorptive part*.

9.3.1 Proof of Unitarity in LSZ Formalism

As mentioned before, let us consider two species of particles, one of mass m (nucleon), associated with $\psi(x)$, and another of mass $\mu < 2m$ (meson), associated with the field $\phi(x)$. The momenta of nucleons are denoted by $p, p'...$, while those of the mesons are denoted by $q, q'...$. Now, there are two distinct types of Klein-Gordon operators which we shall denote as $\mathcal{K}_{x,\mu}, \mathcal{K}_{x,m}$. Consider the process

$$N(p) + M(q) \rightarrow N(p') + M(q') \quad (9.22)$$

The proof consists of directly evaluating the l.h.s. of Eq. (9.21) from the LSZ reduction formulae. The manipulations are much easier in the two space-time integration representations. Therefore, we start with such a representation when the two mesons have been “reduced” (as it happens all the time, the metric conventions of Lehmann and Barton are opposite to each other; at this stage with only spinless particles, it is easy to pass from one to another. Every invariant $a \cdot b$ changes sign as we go from one to another. The Klein-Gordon operator in Lehmann’s metric is $(\square - m^2)$). The relevant representation, in Barton’s metric, is

$$-i(2\pi)^4 \bar{T}(p'q'; pq) = - \int \int dx dy e^{iq' \cdot x - iq \cdot y} \mathcal{K}_{x,\mu} \mathcal{K}_{y,\mu} \langle p' | [\phi(x), \phi(y)]_{ret} | p \rangle \quad (9.23)$$

Likewise,

$$-i(2\pi)^4 \bar{T}(pq; p'q') = - \int \int dx dy e^{iq' \cdot x - iq \cdot y} \mathcal{K}_{x,\mu} \mathcal{K}_{y,\mu} \langle p | [\phi(x), \phi(y)]_{ret} | p' \rangle \quad (9.24)$$

Taking the complex conjugate of this equation,

$$+ i(2\pi)^4 \bar{T}^*(pq; p' q') = - \int \int dx dy e^{-iq \cdot x + iq' \cdot y} \mathcal{K}_{x,\mu} \mathcal{K}_{y,\mu} \langle p' | [\phi(x), \phi(y)]_{ret}^\dagger | p \rangle \quad (9.25)$$

On noting $[\phi(x), \phi(y)]^\dagger = [\phi(y), \phi(x)]$, this can be further rewritten as

$$+ i(2\pi)^4 \bar{T}^*(pq; p' q') = - \int \int dx dy e^{-iq \cdot x + iq' \cdot y} \mathcal{K}_{x,\mu} \mathcal{K}_{y,\mu} \theta(x_0 - y_0) \langle p' | [\phi(y), \phi(x)] | p \rangle \quad (9.26)$$

We rewrite it further by interchanging x, y on the r.h.s.:

$$+ i(2\pi)^4 \bar{T}^*(pq; p' q') = - \int \int dx dy e^{-iq \cdot y + iq' \cdot x} \mathcal{K}_{x,\mu} \mathcal{K}_{y,\mu} \theta(y_0 - x_0) \langle p' | [\phi(x), \phi(y)] | p \rangle \quad (9.27)$$

Adding Eqs. (9.23)–(9.27) yields

$$\begin{aligned} & i(2\pi)^4 \{ \bar{T}^*(pq; p' q') - \bar{T}(p' q'; pq) \} \\ &= - \int \int dx dy e^{iq' \cdot x - iq \cdot y} \mathcal{K}_{x,\mu} \mathcal{K}_{y,\mu} \langle p' | [\phi(x), \phi(y)] | p \rangle \end{aligned} \quad (9.28)$$

Remarkably, the θ -function has disappeared on using the identity $\theta(x) + \theta(-x) = 1$! The $\text{Im}T$ (in the sense described), only involves the *unretarded commutators*! This will have several deep consequences, as we shall see. First, let us recast this in terms of a single x -integration representation. On noting that $\mathcal{K}_{x,\mu} \phi(x) = j(x)$, the r.h.s. becomes

$$- \int \int dx dy e^{iq' \cdot x - iq \cdot y} \langle p' | [j(x), j(y)] | p \rangle \quad (9.29)$$

This is simplified on using displacement operators $e^{iP \cdot a}$ with the action

$$e^{iP \cdot a} O(x) e^{-iP \cdot a} = O(x + a) \quad (9.30)$$

There is some discrepancy between these operators as introduced by Gasiorowicz (Eqs. (1.36) and (1.23) of [12]) and Barton (see Eq. (2.2) on p. 8 of [11]), notwithstanding the fact their metrics are the same). In fact, all the books have thoroughly messed up introducing coordinates and momenta. This may perhaps be because of the feeling that the distinction between covariant and contravariant vectors is not important in flat space. The coordinates should be x^μ , while momenta should be p_μ , but in the end, one should understand in terms of $t, \mathbf{x}, E, \mathbf{p}$ etc. On using the displacement operator as mentioned above,

$$- \int \int dx dy e^{i(q' \cdot x - q \cdot y + \frac{(p' - p)}{2} \cdot (x + y))} \langle p' | [j(\frac{x - y}{2}), j(\frac{y - x}{2})] | p \rangle \quad (9.31)$$

On introducing $x' = x - y, y' = x + y$, and performing the integration over y' , it is straightforward to see that this becomes

$$- (2\pi)^4 \delta^{(4)}(p' + q' - p - q) \int dx' e^{i \frac{(q + q')}{2} \cdot x'} \langle p' | [j(\frac{x'}{2}), j(-\frac{x'}{2})] | p \rangle \quad (9.32)$$

If we had chosen Barton's displacement operators, the overall energy-momentum delta function would not even have come out right! Peeling off the $(2\pi)^4 \delta^{(4)}(p' + q' - p - q)$ from both sides, we write $\text{Im}T$ as

$$i \langle T^*(pq; p' q') - T(p' q'; pq) \rangle = - \int dz e^{i \frac{(q+q')}{2} \cdot z} \langle p' | j(\frac{z}{2}), j(-\frac{z}{2}) | p \rangle \quad (9.33)$$

The next step is to show that this is exactly the same as the unitarity condition. Let us first analyse the $\langle p' | j(\frac{z}{2}) j(-\frac{z}{2}) | p \rangle$ term of the commutator. Introducing a complete set of states, and using the displacement operators, it is easy to see that this is

$$\langle p' | j(\frac{z}{2}) j(-\frac{z}{2}) | p \rangle = \sum_n e^{i \frac{(p'+p-2p_n)}{2} \cdot z} \langle p' | j(0) | n \rangle \langle n | j(0) | p \rangle \quad (9.34)$$

The z -integration gives $(2\pi)^4 \delta^{(4)}(p_n - \frac{p+q+p'+q'}{2})$, which is the same as $(2\pi)^4 \delta^{(4)}(p_n - p - q)$. This means the intermediate states that contribute are precisely those with the same initial (final) energy-momentum (it is worth pointing out again that only the displacement operators of Eq. (9.30) can lead to this). This is of course a necessary condition for the r.h.s. to be the same as in the unitarity condition. But it is not sufficient and, for that, we have to show that the factors, $\langle p' | j(0) | n \rangle$, $\langle n | j(0) | p \rangle$ are the correct T-matrix elements that enter the unitarity relation.

To prove this, we step back to the very first step in arriving at the LSZ reduction formulae; when one of the incoming particles was reduced, this led to Eq. (8.13), which we reproduce here for continuity:

$$\langle \alpha' \beta', out | \alpha \beta, in \rangle = \langle \alpha' \beta', in | \alpha \beta, in \rangle + i \int dy e^{-ip_\beta \cdot y} \mathcal{K}_y \langle \alpha' \beta', out | \phi(y) | \alpha \rangle \quad (9.35)$$

This can be straight away taken over to the present context (modulo the trivial contribution, and after using $\mathcal{K}_{y,\mu} \phi(y) = j(y)$) as

$$\begin{aligned} \langle n, out | pq, in \rangle &= -i(2\pi)^4 \delta^{(4)}(p_n - p - q) T(n; pq) \\ &= i \int dy e^{-q \cdot y} \langle n, out | j(y) | p \rangle \end{aligned} \quad (9.36)$$

It is instructive to write the r.h.s. of this equation as

$$i \langle n, out | j(q) | p \rangle \quad (9.37)$$

which brings out the role of $j(q)$ in creating a meson of momentum q , and also explains the occurrence of the $\delta^{(4)}(p_n - p - q)$. Returning to the r.h.s. of Eq. (9.36),

$$\begin{aligned} i \int dy e^{-iq \cdot y} \langle n, out | j(y) | p \rangle &= i \int dy e^{-i(q+p-p_n) \cdot y} \langle n, out | j(0) | p \rangle \\ &= i(2\pi)^4 \delta^{(4)}(p_n - p - q) \langle n, out | j(0) | p \rangle \end{aligned} \quad (9.38)$$

Which means

$$\langle n, out | j(0) | p \rangle = -T(n; pq) \quad (9.39)$$

thus proving the sufficiency for the proof of unitarity in the LSZ scheme. Before concluding, we have to address the second term in the commutator. In complete parallel to Eq. (9.34),

$$\langle p' | j(-\frac{z}{2}) j(\frac{z}{2}) | p \rangle = \sum_n e^{-i \frac{(p' + p - 2pn)}{2} \cdot z} \langle p' | j(0) | n \rangle \langle n | j(0) | p \rangle \quad (9.40)$$

The z -integration after including the $e^{i \frac{(q+q')}{2} \cdot z}$ factor now yields

$$\delta^{(4)}(p_n - \frac{p - q' + p' - q}{2}) \quad (9.41)$$

This can never be satisfied for s-channel physical processes and the proof of s-channel unitarity in the LSZ scheme is complete. But the scheme is capable of more. The discarded part of the commutator while being irrelevant for s-channel physical processes, has a deeper meaning of its own. The delta-function which could not be satisfied by s-channel kinematics, can, nevertheless, be satisfied by t-channel physical processes. Thus, the LSZ scheme incorporates unitarity in *all* the physical channels.

References

1. K. Symanzik, Phys. Rev. **105**, 743 (1957)
2. R. Capps, G. Takeda, Phys. Rev. **103**, 1877 (1956)
3. A. Salam, Nuovo Cimento **3**, 424 (1956)
4. G.F. Chew, M.L. Goldberger, F.E. Low, Y. Nambu, Phys. Rev. **106**, 1337 (1957)
5. G.F. Chew, M.L. Goldberger, F.E. Low, Y. Nambu, Phys. Rev. **106**, 1345 (1957)
6. A. Martin, *The Rigorous Analyticity-Unitarity Program and its Successes*, in Ringberg Symposium on QFT (1998)
7. N.N. Khuri, in *Summer School in Elementary Particle Physics*. Brookhaven National Lab., July 22–Aug 29 1969
8. H. Lehmann, Nuovo Cimento 10 No 4, p. 579 (1958)
9. H. Lehmann, Supp. Nuovo Cimento **14**, 153 (1959)
10. G. Sommer, Fortschritte Phys. **18**, 577 (1970)
11. G. Barton, *Introduction to Advanced Field Theory*. Interscience Tracts on Physics and Astronomy, vol. 22 (1963)
12. S. Gasiorowicz, *Elementary Particle Physics*. Wiley

10.1 Lehmann Ellipses

Now, we turn to an entirely different use of the LSZ scheme. This is the issue of the analytic properties of the scattering amplitude as a function of angle, or equivalently, the momentum transfer, for fixed energy. This should be contrasted with the analytic properties of the scattering amplitude as a function of energy for fixed momentum transfer. We shall return to the latter when we take up *Dispersion Relations*.

Without performing any calculations, one can get an idea about how to extend the domain of analyticity in scattering angle (momentum transfer). The physical region is characterized by *real* scattering angles with the variable $z = \cos \theta$ also real and in the range $-1 \leq z \leq 1$. The simplest way to analytically extend this domain is to let the scattering angle, say, θ , and hence z , become *complex*. In other words, $\theta = \theta_r + i\theta_I$. This leads to

$$z = \cos(\theta_r + i\theta_I) = \cosh(\theta_I) \cos(\theta_r) - i \sinh(\theta_I) \sin(\theta_r) \quad (10.1)$$

In other words, $\text{Re}z$ and $\text{Im}z$ are traced by the curve

$$\left(\frac{\text{Re}z}{\cosh(\theta_I)}\right)^2 + \left(\frac{\text{Im}z}{\sinh(\theta_I)}\right)^2 = 1 \quad (10.2)$$

This is nothing but an *ellipse*, with semi-major axis, usually denoted by a , given by $a = \cosh(\theta_I)$, and likewise, the semi-minor axis b given by $b = \sinh(\theta_I)$. The foci of the ellipse are given by $\pm c$ with $c = \sqrt{(a^2 - b^2)}$. Therefore, the foci of our ellipse are at ± 1 , which are just the extremities of the unextended physical region. The domain of analyticity has now been extended to the entire ellipse. Another relation that will prove to be useful later is

$$a + b = e^{\theta_I} \quad (10.3)$$

giving a direct relation between θ_I and the sum of the semi-major and semi-minor axes.

Lehmann's great work lay in actually determining θ_I from non-perturbative QFT, essentially using the LSZ formalism along with the powerful Jost-Lehmann-Dyson theorem. All the important features like the domain being an ellipse with foci at ± 1 stay unchanged by his rather sophisticated analysis. Of greatest significance is the fact that Lehmann could determine the dependence of θ_I on energies, particle masses as well as *inelastic thresholds* and these will be seen to be crucial for several important applications like fixed-t dispersion relations, asymptotic behaviour of partial wave amplitudes (which in turn have important bearing on analytic extension to complex angular momentum plane, Regge Poles, etc.).

Following Lehmann [1], we focus on spin-less nucleons of mass m and mesons of mass μ . The process we consider is the scattering of incoming nucleons of momentum p , mesons of momentum k , into outgoing nucleons of momentum p' and mesons of momentum k' .

As already stressed, there are many LSZ reductions possible for the same scattering amplitude, and the one identified by Lehmann (Eq. (2) of [1]) reads

$$T = \int dx e^{i\frac{(k'-p')}{2} \cdot x} \langle 0 | R' \{ \phi(\frac{x}{2}) \psi(-\frac{x}{2}) \} | p k, in \rangle \quad (10.4)$$

The R' -products are as defined in Eq. (9.5) where they were defined for two ϕ -fields. Here, we need such an R' -product for a ϕ -field with a ψ -field, and is given by

$$R' \{ \phi(x) \psi(y) \} = -i \mathcal{K}_{x,\mu} \mathcal{K}_{y,m} \theta(x_0 - y_0) [\phi(x), \psi(y)] \quad (10.5)$$

In this way of reduction, both the final nucleon and meson have been reduced. The expressions have already been brought to the single space-time integration form. Here too, the imaginary part of T can be shown to be given in terms of *unretarded commutators*! For example, we explicitly showed this by deriving Eq. (9.28) starting from Eq. (9.23). Same techniques can be used for any T -matrix element of choice. Denoting generic T -matrix elements as Fourier-Transforms $F_{R'}(q)$ (Lehmann calls these $f_R(q)$ which is somewhat misleading as the T -matrix elements are not FT of the matrix-elements of his R -products of Eq. (9.6)), Lehmann writes down the representation

$$F_{R'}(q) = -\frac{1}{2\pi} \int dq'_0 \frac{F(q'_0, \mathbf{q})}{q'_0 - q_0} \quad \text{Im } q_0 \geq 0 \quad (10.6)$$

where $F(q)$ is the FT of the corresponding *unretarded* commutator. For the case at hand,

$$F(q) = \int dx e^{iq \cdot x} \langle 0 | [j(\frac{x}{2}), f(-\frac{x}{2})] | p + k, \gamma \rangle \quad (10.7)$$

Here, $j(x) = \mathcal{K}_{x,\mu} \phi(x)$, $f(x) = \mathcal{K}_{x,m}$ are, respectively, the sources for mesons and nucleons. $|p + k, \gamma\rangle$ collectively denotes all states with total energy-momentum

$p + k$. Also, for the case under analysis, $q = k' - p'$, but we shall continue for a general q .

It is instructive to dwell a bit on the meaning of this integral equation. Lehmann, in his lecture notes [2] alludes to this as being in a *formal manner*. Recalling that the FT of the θ -function is (see Appendix C of [3])

$$\theta(t) = \frac{-1}{2\pi i} \int_{-\infty}^{\infty} \frac{d\omega}{\omega + i\epsilon} e^{-i\omega t} \quad (10.8)$$

one may be tempted to interpret Eq. (10.6) as a *convolution*. But the R' -product is not a simple product of a $\theta(x_0 - y_0)$ with another function of $(x - y)$, because the Klein-Gordon operators act both on the matrix elements of a commutator, and, the θ -function. The other interpretation of Eq. (10.6) is that it embodies the connection between $\text{Im } F_{R'}$ (in the sense of discontinuities rather than as imaginary parts of complex functions) and F , on recognizing the well-known identity

$$\frac{1}{\omega - i\epsilon} = P \frac{1}{\omega} + i\pi \delta(\omega) \quad (10.9)$$

Lehmann's rather brilliant strategy is to use this integral equation to determine $F_{R'}(q)$ from a knowledge of $F(q)$, and that the constraints of *microcausality* on F are more direct than on their constraints on $F_{R'}(q)$.

10.1.1 Jost-Lehmann-Dyson Theorem

On introducing a complete set of states between $j(\frac{x}{2})$ and $f(-\frac{x}{2})$, it is easy to see that $F(q)$ vanishes, unless (i) $\frac{p_0+k_0}{2} + q_0 \geq 0$, and, $(\frac{p+k}{2} + q)^2 \geq m_1^2$, or, (ii) $\frac{p_0+k_0}{2} - q_0 \geq 0$, and, $(\frac{p+k}{2} - q)^2 \geq m_2^2$. Here, m_1, m_2 are the *lowest* masses of the intermediate states contributing to the two terms in the commutator. Specifically, for the $j(\frac{x}{2})f(-\frac{x}{2})$ term, the lowest mass for the intermediate states is m_1 , and likewise m_2 for the other term $f(-\frac{x}{2})j(\frac{x}{2})$ term. Then the above conditions are simply the conditions of the positivity of energy and that masses of the other intermediate states have to be higher.

What can one say about m_1 and m_2 ? To that end, let us recall the single-particle stability conditions discussed in Chap. 6. In particular, the Eq. (6.35) tells us that $\langle 0|j|k\rangle$ and $\langle 0|f|p\rangle$ vanish for all k, p . Therefore, the intermediate states that contribute in the first case are *multimeson states*. If the mesons are *scalar*, the lowest mass intermediate states are *two meson states* with $m_1 = 2\mu$. If, on the other hand, the mesons are *pseudoscalar*, the lowest mass intermediate states are *three meson states* with $m_1 = 3\mu$. Lehmann treats this latter case. Irrespective of whether mesons are scalar or pseudoscalar, the lowest mass intermediate state for the other case is one with a single nucleon and a meson, with $m_2 = m + \mu$.

Jost and Lehmann [4] first solved the problem of finding functions satisfying the above conditions, though only for the equal mass cases of $m = \mu$. Dyson [5]

extended the proof to cover the unequal mass cases also. As the masses will be seen to play crucial roles in determining domains of analyticity, we, following Lehmann, shall treat the unequal mass case.

We refer the readers to [4,5] for the proofs of the theorem. We shall state their final result. It is necessary and sufficient for an $F(q)$ to satisfy the two sets of two conditions if it can be represented as

$$F(q) = \int d^4u \int_0^\infty d\kappa^2 \epsilon(q_0 - u_0) \delta((q - u)^2 - \chi^2) \Phi(u, \kappa^2) \quad (10.10)$$

Here, $\epsilon(x) = 1$ for $x > 0$, and $\epsilon(x) = -1$ for $x < 0$. Rather remarkably, $\Phi(u, \kappa^2)$ is *any* function if both the four vectors $\frac{(p+k)}{2} + u$, and, $\frac{(p+k)}{2} - u$ lie in the *forward light-cone*. Recall that a vector k lies in the forward light-cone if $k_0 > 0$ and $k^2 > 0$. In addition to the forward light-cone restrictions, κ has to satisfy a non-trivial constraint too:

$$\kappa \geq \max\{0; m_1 - \sqrt{(\frac{(p+k)}{2} + u)^2}; m_2 - \sqrt{(\frac{(p+k)}{2} - u)^2}\} \quad (10.11)$$

Outside this region, $\Phi(u, \kappa^2)$ vanishes. The reader is warned about some notational inconsistencies in [1]; he uses “ u ” both for the four-vector u , as well as for the magnitude of $\mathbf{u}$. In the above, u refers to the four-vector (as is also clear from the context!). The strength of the Jost-Lehmann-Dyson theorem is that it can be so succesfully applied to problems of analyticity of scattering amplitudes without knowing anything more about $\Phi(u, \kappa^2)$. The weakness of the theorem is, at least till now, that it does not say anything about what Φ may actually be. Clearly, it must depend on the operators and states that go into $F(q)$ like j, f, γ . Stated differently, a way to invert Eq. (10.10) to obtain Φ would be desirable, as its explicit form may yield even more powerful ways for determining the analytic behaviours.

Substituting Eq. (10.10) in Eq. (10.6), one gets

$$F_{R'}(q) = -\frac{1}{2\pi} \int \int d^4u d\kappa^2 \Phi(u, \kappa^2) \int \frac{dq'_0}{q'_0 - q_0} \epsilon(q'_0 - u_0) \delta((q'_0 - u_0)^2 - (\mathbf{q} - \mathbf{u})^2 - \kappa^2) \quad (10.12)$$

Had one integrated over κ^2 first, the q -dependence of $F_{R'}(q)$ would have got buried in the Φ . That is why Lehmann has chosen this order of integration. The integration over q'_0 can be performed on first recognizing that the delta-function is a quadratic in q'_0 , i.e. $\delta((q'_0 - u_0)^2 - (\mathbf{q} - \mathbf{u})^2 - \kappa^2)$. The roots of this quadratic form are at $q'_0 = u_0 \pm \sqrt{(\mathbf{q} - \mathbf{u})^2 + \kappa^2}$. One uses the well-known property of delta-functions:

$$\delta(f(x)) = \sum_i \frac{\delta(x - x_i)}{|f'(x_i)|} \quad (10.13)$$

where x_i are the roots of $f(x) = 0$ and $f'(x_i)$ are the first derivatives of $f(x)$ evaluated at $x = x_i$. It is straightforward to put all the ingredients together to arrive at the elegant result:

$$F_{R'}(q) = \frac{-1}{2\pi} \int d^4u \int d\kappa^2 \frac{\Phi(u, \kappa^2)}{(q-u)^2 - \kappa^2} \quad (10.14)$$

Every factor in the Jost-Lehmann-Dyson representation is crucial. For example, if the $\epsilon(q_0 - u_0)$ were not there, one would not even get a Lorentz-covariant answer. The Lorentz-covariance of the final result is also noteworthy as the integral equation (10.6) itself lacks manifest covariance. The main power and usefulness of Eq. (10.14) lies in the fact that the entire q -dependence has been isolated in the denominator.

We remark in passing that Φ in principle has dependence on $p + k$ and the states γ . Therefore, the Lehmann result for the T -matrix should be written as

$$T = \frac{1}{2\pi} = \int \int d^4u d\kappa^2 \frac{\Phi(u, \kappa^2, p, k, \gamma)}{(\frac{(k'-p')}{2} - u)^2 - \kappa^2} \quad (10.15)$$

Though this looks very much like a one-variable *Dispersion Relation*, it is very different. The Lorentz invariance of this representation means that we can evaluate the denominator in *any Lorentz frame*. A naturally convenient choice is the so-called *centre of mass system* (it should however be remembered that the system is fully relativistic). In this frame, the initial and final four momenta are

$$p = (E_p, \mathbf{p}), k = (\omega_p, -\mathbf{p}); p' = (E_{p'}, \mathbf{p}'), k' = (\omega_{p'}, -\mathbf{p}') \quad (10.16)$$

where $E_q = \sqrt{\mathbf{q}^2 + m^2}$, $\omega_q = \sqrt{\mathbf{q}^2 + \mu^2}$. In this case where the masses of nucleons and mesons are the same in both initial and final state, it is a standard exercise in relativistic kinematics to show that actually $\mathbf{p}^2 = \mathbf{p}'^2$. It then follows that W , the total centre of mass energy, and K , the common magnitudes of all the momenta are related according to

$$K^2 = \frac{[W^2 - (m + \mu)^2][W^2 - (m - \mu)^2]}{4W^2} \quad (10.17)$$

In the centre of mass frame, the restrictions of the Jost-Lehmann-Dyson theorem become

$$\begin{aligned} 0 \leq u \leq \frac{W}{2}; -\frac{W}{2} + u \leq u_0 \leq \frac{W}{2} - u \\ \kappa \geq \text{Max}\{0, m_1 - \sqrt{(\frac{W}{2} + u_0)^2 - u^2}; m_2 - \sqrt{(\frac{W}{2} - u_0)^2 - u^2}\} \end{aligned} \quad (10.18)$$

The Mandelstam variables are now $s = (p + k)^2 = (p' + k')^2 = W^2$, $t = (p - p')^2 = (k' - k)^2$, $u = (p - k')^2 = (p' - k)^2$, satisfying the constraint $s + t + u = 2m^2 + 2\mu^2$. The scattering angle θ is related to the t-invariant by

$$t = -4 \Delta^2 = -2 K^2 (1 - \cos \theta) \quad (10.19)$$

The s-channel physical region is given by $s \geq (m + \mu)^2$; $-4K^2 \leq t \leq 0$. The common magnitude of all the momenta gives rise to these useful formulae:

$$E_p = E_{p'} = \frac{(W^2 + m^2 - \mu^2)}{2W} \quad \omega_p = \omega_{p'} = \frac{(W^2 + \mu^2 - m^2)}{2W} \quad (10.20)$$

There are three 3-vectors in the problem: $(\mathbf{p}, \mathbf{p}', \mathbf{u})$. Lehmann introduces *polar coordinates* u, β, α for $\mathbf{u}$, with the ranges $0 \leq \alpha \leq \pi, 0 \leq \beta \leq 2\pi$. Then

$$\int d^4u \int d\kappa^2 \rightarrow \int_0^\infty d\kappa^2 \int_{-\infty}^\infty du_0 \int_0^\infty u^2 du \int_0^{2\pi} d\alpha \int_0^\pi \sin \beta d\beta \quad (10.21)$$

Lehmann seems to have fixed the various polar angles according to

$$\mathbf{p} = K \hat{x}; \mathbf{p}' = K(\cos \theta \hat{x} + \sin \theta \hat{y}); \mathbf{u} = (\cos \beta \hat{z} - \sin \beta \cos \alpha \hat{x} - \sin \beta \sin \alpha \hat{y}) \quad (10.22)$$

It is then straightforward to show that the denominator in Eq. (10.15) is

$$-K^2 - u^2 - \kappa^2 + (u_0 + \frac{(m^2 - \mu^2)}{2W})^2 + 2Ku \sin \beta \cos(\theta - \alpha) \equiv -2Ku \sin \beta (x - \cos(\theta - \alpha)) \quad (10.23)$$

Consequently, Lehmann rewrites Eq. (10.15) as (details can be found in [1])

$$T(W, \cos \theta) = \int_{x_0(W)}^\infty dx \int_0^{2\pi} d\alpha \frac{\bar{\Phi}(x, \alpha, W)}{x - \cos(\theta - \alpha)} \quad (10.24)$$

The minimum value of x when u_0, u, κ are restricted to their ranges as in Eq. (10.18) is $x_0(W)$:

$$x_0(W) = \text{Min}\left\{\frac{K^2 + u^2 + \kappa^2 - (u_0 + (\frac{m^2 - \mu^2}{2W})^2)}{2Ku}\right\} \quad (10.25)$$

This minimization has to be done respecting the ranges of Eq. (10.18), and is not that straightforward. We just quote the explicit expression given by Lehmann

$$x_0(W) = \left\{1 + \frac{(m_1^2 - \mu^2)(m_2^2 - m^2)}{K^2(W^2 - (m_1 - m_2)^2)}\right\}^{\frac{1}{2}} \quad (10.26)$$

It is quite amazing that Lehmann was able to obtain an explicit formula for this, as it is going to play many important roles later on.

Several aspects of this formula are worth emphasizing. The foremost is that $x_0(W) \geq 1$. This means that for real scattering angles, the denominator can never vanish. This has the significance that $T(W, \cos \theta)$ is an *analytic function* of $\cos \theta$. The analyticity can break down when the denominator vanishes. That can certainly happen when θ is extended to complex values.

There are several ways to explore this possibility. Considering complex $\theta = \theta_r + i\theta_I$, we see that the denominator can vanish when $\theta_r = \alpha$, and, $\cosh \theta_I = x$. With

$x_0(W)$ being the smallest value of x , it follows that the scattering angle can be analytically continued till at least this value of θ_I . We also saw, from rather general considerations, that extending scattering angle to complex values is tantamount to the analyticity domain within an ellipse with semimajor axis $\cosh \theta_I$, semi-minor axis $\sinh \theta_I$, and foci at ± 1 . Lehmann's theory shows that the semi-major axis is $x_0(W)$, the semi-minor axis is $\sqrt{x_0^2 - 1}$. In terms of $x_0(W)$, θ_I is also given by

$$e^{\theta_I} = x_0(W) + \sqrt{x_0^2 - 1} \quad (10.27)$$

Lehmann arrived at these in a slightly different way. He asked where the denominator $x - \cos(\theta - \alpha)$ would vanish for complex θ (because $x \geq 1$, it can never vanish for real θ). It is easy to see that the vanishing of the denominator leads to the following quadratic equation for $z = \cos \theta$ (complex). It should be noted that standard trigonometric relations like $\cos^2 \theta + \sin^2 \theta = 1$ continue to hold even for complex θ !

$$z^2 - 2xz \cos \alpha + (x^2 - \sin^2 \alpha) = 0 \quad (10.28)$$

The roots of this quadratic equation are,

$$z = x \cos \alpha \pm i \sqrt{x^2 - 1} \sin \alpha \quad (10.29)$$

We again recognize the Lehmann Ellipse:

$$\left(\frac{\operatorname{Re} z}{x}\right)^2 + \left(\frac{\operatorname{Im} z}{\sqrt{x^2 - 1}}\right)^2 = 1 \quad (10.30)$$

This ellipse is called the *Small Lehmann Ellipse*. So far, these results imply that both $\operatorname{Re} T$ and $\operatorname{Im} T$ are analytic functions of $\cos \theta$, *regular* inside the small Lehmann ellipse. The fact that Lehmann gives $x_0(W)$ explicitly as a function of W , K , and the masses m , μ as well as the thresholds m_1 , m_2 has a lot of implications for analyticity of scattering amplitudes. But before taking them up, let us discuss the second part of Lehmann's work where he shows that $\operatorname{Im} T$ is actually regular in a much larger ellipse called the *Large Lehmann Ellipse*.

This is based on exploiting the unitarity constraints on T . We refer the reader to the many algebraic details, but shall only present the essentials. As discussed before, the unitarity relation is a *non-linear* relation between T 's:

$$i(T^*(pq; p'q') - T(p'q'; pq)) = -(2\pi)^4 \sum_n \delta^{(4)}(p_n - p - q) T^*(n; p'q') T(n; pq) \quad (10.31)$$

Now, Lehmann's idea is to use the Jost-Lehmann-Dyson representations of the form of Eq. (10.15) for each of the T 's on the r.h.s. Now we shall have the variables

$u_1, u_2, \kappa_1, \kappa_2$, or, in terms of polar decompositions $u_1, u_2, u_{01}, u_{02}, \alpha, \beta_1, \beta_2, \chi$. The only detail that will be relevant is that

$$\begin{aligned} \operatorname{Im} T &\simeq \int_0^{2\pi} \frac{1}{[x_1 - \cos(\theta - \chi)]} \cdot \frac{1}{[x_2 - \cos(\chi - \alpha)]} \\ &= 2\pi \frac{\left(\frac{x_1}{\sqrt{x_1^2 - 1}}\right) + \left(\frac{x_2}{\sqrt{x_2^2 - 1}}\right)}{(x_1 x_2 + \sqrt{x_1^2 - 1} \sqrt{x_2^2 - 1} - \cos(\theta - \alpha))} \end{aligned} \quad (10.32)$$

Here, x_1, x_2 are the analogs of the x introduced earlier for the two different sets of parameters associated with the two T -matrices in the product. As before, the minimum value of $y = x_1 x_2 + \sqrt{x_1^2 - 1} \sqrt{x_2^2 - 1}$ has to be determined. Here, minimization is with respect to the parameters of the Jost-Lehmann-Dyson representation. The result is straightforward: $x_1 = x_2 = x_0(W)$, with $x_0(W)$ as given before. Therefore, $y_0 = 2x_0^2 - 1$. The resultant expression for $\operatorname{Im} T$ is

$$2 \operatorname{Im} T = \int_{2x_0^2 - 1}^{\infty} dy \int_0^{2\pi} d\alpha \frac{\tilde{\Phi}(y, \cos \alpha, W)}{y - \cos(\theta - \alpha)} \quad (10.33)$$

As before, this means that $\operatorname{Im} T$ is an analytic function of $\cos \theta$ regular in a much larger domain than $\operatorname{Re} T$, called the *Large Lehmann Ellipse*. It's semi-major axis is given by $y_0 = 2x_0^2 - 1$, and its semi-minor axis by $\sqrt{y_0^2 - 1} = 2x_0 \sqrt{x_0^2 - 1}$. The foci of this larger ellipse are also at ± 1 . It is instructive to calculate $\theta_{I,L}$ for this larger ellipse and compare it to the $\theta_{I,S}$ for the smaller ellipse which was given by $e^{\theta_I} = x_0 + \sqrt{x_0^2 - 1}$.

$$e^{\theta_{I,L}} = y_0 + \sqrt{y_0^2 - 1} = (x_0 + \sqrt{x_0^2 - 1})^2 = e^{2\theta_{I,S}} \quad (10.34)$$

Thus, $\theta_{I,L}$ is exactly twice as large as $\theta_{I,S}$! This is a considerable enlargement of the analyticity domain.

The formulae for $x_0(W)$, $y_0(W)$ explicitly show their dependence on K, W, m , as well as μ, m_1, m_2 . With their help, one can investigate analyticity domains as these are varied. For example, at threshold, when $W \rightarrow m + \mu$, $K \rightarrow 0$, both $x_0, y_0 \rightarrow \infty$ and the analyticity domain in z is the largest. In contrast, at very high energies, both $x_0(W), y_0(W) \rightarrow 1$ and the analyticity domains in z shrink to their minimum.

But one is often more interested in domains of regularity in t -plane, rather than in z . The latter is of greater interest while investigating partial wave analysis. Since t is related to z by $t = -2K^2(1 - z)$, the t -analyticity domain can be substantial even if the region of z -analyticity is reduced, as happens at very high energies.

Our last remark pertains to the intriguing possibilities of analyticities in *masses*. In the entire derivation of the ellipses, no where was it necessary for $p^2 = p'^2 = m^2$ or $k^2 = k'^2 = \mu^2$. As Lehmann does, one can take $k^2 = k'^2 = \zeta$; this would

of course change various kinematic expressions for W , K etc. One can then think of analytically continue in ζ , with physical amplitudes as their boundary values at $\zeta = \mu^2$. This line of thinking will prove to be very useful in establishing fixed- t dispersion relations, where analyticity is sought for regions of fixed t , but variable s , whereas here we investigated the problem of fixed s but variable t .

References

1. H. Lehmann, Nuovo Cimento 10 No. 4, p. 579 (1958)
2. H. Lehmann, Supp. Nuovo Cimento **14**, 153 (1959)
3. J.D. Bjorken, S. Drell, *Relativistic Quantum Fields*. McGraw Hill Publishers
4. R. Jost, H. Lehmann, Nuovo Cimento **5**, 1598 (1957)
5. F.J. Dyson, Phys. Rev. **110**, 1460 (1958)



11.1 Introduction

The Dispersion Relations are extremely important to the narrative of this book. It is through their intermediary that the general objectives of the S-matrix programme, namely, to understand strong interactions on the basis of optimal amount of their analyticity properties, get linked in a precise manner to the notion of strings. These analyticity properties could only be understood in terms of Relativistic Quantum Field Theories. The general framework of Lehmann, Jost, Symanzik and Zimmermann (LSZ) was crucial for this.

It is the *Superconvergence Relations* arising from Dispersion Relations that led to the notion of *Duality*. This rather unusual and deep concept is what clearly distinguished strong interactions from, say, Quantum Electrodynamics or the Electroweak unified field theories. How Duality eventually led to the formulation of the string description of hadronic physics will be dealt with in great detail in what follows.

Apart from this central theoretical role played by the Dispersion Relations, they gained prominence as concrete realizations of the S-matrix programme in the sense of explicitly demonstrating the analyticity properties that had been conjectured for scattering amplitudes, and, of providing a link with observations. Many aspects that lay beyond the reach of the perturbative quantum field theories could now be analysed, even quantitatively at times.

From our extensive discussion of the Kramers-Krönig dispersion relations in optics, we recall the following salient aspects: emergence of the analyticity in the upper half frequency (complex) plane as a direct consequence of causal behaviour of physical systems, i.e. the principle of “no output before any input”, the emergence of a direct relationship between the real part of a scattering amplitude and its imaginary part, known in mathematical literature as Hilbert Transforms, and finally, expressing the imaginary part of the amplitude in terms of *directly measurable quantities*. Additionally, the boundedness for large frequencies, existence of special relations

like the symmetry-equation, are also very important. It is important to stress that all these considerations need not be restricted to scattering amplitudes only, they can be applied as long as there is a *linear* relationship between input and output. For example, an electrical circuit characterized by *complex impedance* can also be analysed by dispersion relations; the incoming current would act like an input while the resulting potential would be the output. Actually, in addition to linearity, the system must possess *time translation* invariance.

It was Krönig who made the very important suggestion of augmenting the S-matrix programme, as originally envisaged by Heisenberg (and discussed extensively in Chap. 4), by the requirements of causality and the resultant analyticity. This was followed by heightened activity towards generalizing Kramers and Krönig's works. Schutzer and Tiomno [1] treated the case of non-relativistic quantum mechanics, van Kampen [2,3] made very careful and incisive analyses of both non-relativistic particles as well as light. He pointed out many subtleties in the careful formulation of what one means by causality. Both works were couched in the language of waves and scatterers and were far from the framework of field theory.

11.2 Toll's Analysis of the Logical Foundations

An important work in this connection is John Toll's thesis [4], the first part of which was published as a paper in [5]. His thesis was titled *Dispersion Relations and its application to problems involving electron pairs*. The first part is of specific interest to us in this chapter, and in it Toll claimed a *rigorous proof of the logical equivalence between strict causality and the validity of Dispersion Relations*. In doing so, he characterizes *strict causality* as the "no output before input", and also as "no signal faster than light in vacuum". The latter is somewhat questionable as causal orderings are meaningful only if one considers the *Orthochronous Lorentz Group* whereas the signal propagation being limited by the speed of light in vacuum is preserved under the full Lorentz group.

While Kramers and Krönig showed how causality leads to analyticity, and consequently the Hilbert transforms relating real and imaginary parts, Toll wishes to see this as *logical equivalence*. This means the converse is also true, i.e. the validity of the dispersion relations should imply the requisite analyticity, and as a consequence, strict causality. By his own admission, Toll is more or less doing exactly what Kramers and Krönig did, but make their discussion more precise. He claims that Krönig in his derivation of the dispersion relations had made additional assumptions about analyticity which Toll's treatment avoids. But this author is not aware of any such, and in any case Toll is not very explicit about it. Toll does not say if Kramers also makes similar assumptions. Therefore, while Toll's work did not make progress towards the derivation of dispersion relations from field theory, it gives a new perspective as well as a mathematically more rigorous founding of Kramers-Krönig derivations.

The central idea of Toll's is that of *Logical Equivalence*. Before giving his seven forms of this logical equivalence, let us briefly describe his formalism. The input $F(t)$ and the output $G(t)$ have the linear relationship

$$G(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dt' T(t - t') F(t') \quad (11.1)$$

The time translation invariance, also called a stationary process, is reflected through the fact that $T(t - t')$ is *time translation invariant*. This has the consequence that if the input is chosen as $F(t + \tau)$ instead of $F(t)$, the output will be $G(t + \tau)$. This is a *convolution* in the time variable. Strict causality in this context would imply $T(t) = 0$ for $t < 0$. By the well-known *Faltung Theorem*, the Fourier transforms $f(\omega)$, $T(\omega)$, $g(\omega)$ of $F(t)$, $T(t)$, $G(t)$ respectively have the simple product relation

$$g(\omega_r) = T(\omega_r) \cdot f(\omega_r) \quad (11.2)$$

Here ω_r is the real part of the complex frequency ω , i.e. $\omega = \omega_r + i\omega_i$. Thus, time translation invariance for this problem leads to the great simplification that there is no mixing of frequencies. Toll calls $T(\omega)$ the *generalized scattering amplitude*. In general, Fourier-Transforms may not exist and one may need to invoke some heavy mathematical machinery like *Distributions*. But if the functions are square-integrable, the existence of their Fourier-Transforms is guaranteed. Absolute squares of inputs and outputs typically being *intensities*, this requirement is well motivated on physical grounds.

An important concept that Toll makes use of is that of a *Causal Transform*. We recall here his characterization of a causal transform (immediately after his Eq. (2.4)): *A causal transform is a square-integrable function of a real variable which can be extended almost everywhere to give a function which is analytic in the upper half of the complex plane and which is of uniformly bounded square integral along any line parallel to and above the real axis*. One may wonder about justifying the label *causal* to what appears to be a rather formal definition! The answer lies in Titchmarsh's clarification that a square-integrable function is zero for all negative values of its argument if and only if its Fourier Transform is a causal transform. Thus, $T(\omega_r)$ is a causal form if and only if $T(t)$ obeys strict causality!

Furthermore, Toll mentions the proof by Titchmarsh on the necessary and sufficient conditions for a causal transform; that is if $\phi(\omega_r)$ is a complex function of the real variable, $\phi(\omega_r) = \phi_r(\omega_r) + i\phi_i(\omega_r)$, the *necessary and sufficient* conditions for $\phi(\omega_r)$ to be a causal transform are

$$\phi_r(\omega_r) = \frac{1}{\pi} P \int_{-\infty}^{\infty} d\nu \frac{\phi_i(\nu)}{\nu - \omega_r} \quad \phi_i(\omega_r) = -\frac{1}{\pi} P \int_{-\infty}^{\infty} d\nu \frac{\phi_r(\nu)}{\nu - \omega_r} \quad (11.3)$$

These are exactly the dispersion relations in unsubtracted form, and before any symmetry requirements have been imposed. Thus we see that causal transforms form a nice logical and conceptual bridge between causality and dispersion relations. Since these are both necessary and sufficient, one can work backwards too. That is, starting from dispersion relations one could have arrived at causality as a necessity! This would have been complementary to the approaches of Kramers and Krönig who started with causality and proved dispersion relations as a consequence.

This is in fact the power of logical equivalence. In the specific context of *causality and dispersion relations*, Toll cites seven statements to be logically equivalent. We shall quote the most pertinent ones for our purposes: (i) strict causality, (ii) analyticity in upper half ω -plane, equivalently, the dispersion relations, (iii) *Integrability criterion for all points in lower half plane*: the function $A(\omega_r)$ satisfies

$$\int_{-\infty}^{\infty} d\nu \frac{A(\nu)}{(\nu - \tilde{\omega})^2} = 0 \quad (11.4)$$

for *all* complex numbers $\tilde{\omega}$ in the lower half plane, iv) *Integrability condition for fixed points*: $A(\omega)$ satisfies

$$\int_{-\infty}^{\infty} d\nu \frac{A(\nu)}{(\nu - \tilde{\omega})^{n+1}} = 0 \quad (11.5)$$

for *one particular fixed complex number* $\tilde{\omega}$ in the lower half plane, and for *all* positive integers n , (v) *Strict causality for exponentially decaying input*: for some complex number $\tilde{\omega}$ on the lower half plane, $A(\nu)/(\nu - \tilde{\omega})$ is a *Causal Transform*.

The last three characterizations are usually not to be found in discussions on causality and dispersion relations. Items (iv) and (v) are particularly interesting. They have the flavour of the so called *Superconvergence Relations* which we shall discuss later, and which play a central role in the overall considerations of this book.

11.3 Dispersion Relations in QFT: General Considerations

Let us consider scattering of particles of mass μ , internal quantum numbers α , represented by scalar fields $\phi_\alpha(x)$ by particles of mass M with internal quantum numbers λ, λ' and scalar fields $\psi_\lambda(x)$. To bring out the essentials without cluttering with technical details, we consider the simplest situation where both species are spin-less. The starting point to this discussion could have been the reduction formulae for S-matrix elements that we have already discussed in great detail in the earlier chapters. But we shall instead follow the treatments initiated by Goldberger [6] which actually pre-dated the works of Lehmann, Symanzik, Zimmermann and Jost. Though both approaches will be seen to be identical, the reader is expected to benefit from this larger coverage.

Goldberger starts with what he calls the *Feynman matrix element* for the scattering process $p + q \rightarrow p' + k$ where k, q are the four-momenta of the particles with mass μ . Likewise, p, p' are the four-momenta of the particles with mass M . The Feynman matrix element, his Eq. (2.4), is

$$\mathcal{F}_{\alpha\beta}(k, q; p', p) = i \int dx \int dy e^{-ik \cdot x + iq \cdot y} \mathcal{K}_x \mathcal{K}_y \langle p', \lambda' | T(\phi_\alpha(x) \phi_\beta(y)) | p, \lambda \rangle \quad (11.6)$$

Here $\mathcal{K}$ are the Klein-Gordon operators for the fields $\phi_\alpha(x)$. This is in complete agreement with what the LSZ reduction formulae would have given when the two

$\phi(x)$ fields were reduced, except for a factor of i . It is puzzling that Goldberger calls this a *Feynman* matrix element. While the LSZ formalism from which this follows is fully *non-perturbative*, whatever Feynman (also Schwinger and Tomonaga) developed in the context of QED and also Nuclear Theory were distinctly *Perturbative*. Goldberger does not cite any work of Feynman to support this nomenclature. The best justification that comes to mind is that the above expressions agree with perturbative field theoretic results, for all orders, in the limit when perturbation theory is valid.

The action of the Klein-Gordon operators on the T -product can be worked out. Under some approximations about the nature of interactions like absence of derivative couplings, etc., considered reasonable in those times, Goldberger obtains (his Eq. (2.5))

$$\begin{aligned} & \mathcal{F}_{\alpha\beta}(k, q; p'\lambda', p\lambda) \\ &= i \int \int dx dy e^{-ik \cdot x + iq \cdot y} \langle p'\lambda' | \{ T(j_\alpha(x) j_\beta(y)) - \delta(x_0 - y_0) [j_\alpha(x), \dot{\phi}_\beta(y)] \} | p\lambda \rangle \end{aligned} \quad (11.7)$$

here, $j_\alpha(x) = \mathcal{K}_x \phi_\alpha(x)$ is the source of the $\phi_\alpha(x)$ field. It is easy to see that if one works out the action of the Klein-Gordon operators in the most general case either in Eq. (11.6), or in the reduction formulae of the LSZ formalism, one will not get just Eq. (11.7). In general, there will be additional terms involving $\delta(x_0 - y_0)$, and even their higher derivatives. Goldberger in [7] explains why he drops some of these terms because of the assumptions made about $j_\alpha(x)$. A justification for the neglect of these terms is also given on the basis that they are not relevant for the analyticity structure of the concerned amplitudes. The reader is encouraged to work these in detail and see if the assumptions for dropping them are reasonable. We will be commenting on these terms as we go along.

It is useful to peel off the inevitable $(2\pi)^4 \delta^{(4)}(p' + k - p - q)$ factor from the scattering amplitude. We have already discussed and explicitly demonstrated how to do this using translational invariance. Goldberger chooses to translate by $-y$ to get

$$\begin{aligned} & F_{\alpha\beta}(k, q; p'\lambda', p\lambda) \\ &= i \int dx e^{-ik \cdot x} \langle p'\lambda' | \{ T(j_\alpha(x) j_\beta(0)) - \delta(x_0) [j_\alpha(x), \dot{\phi}_\beta(0)] \} | p\lambda \rangle \end{aligned} \quad (11.8)$$

where

$$F_{\alpha\beta}(k, q; p'\lambda', p\lambda) \equiv (2\pi)^4 F_{\alpha\beta}(k, q; p'\lambda', p\lambda) \quad (11.9)$$

is just the T -matrix element introduced in the earlier chapters. The integral representation for the T -matrix here looks very different than what we got earlier and what Lehmann used in obtaining the Lehmann ellipses. There, the exponential factor was $e^{-i \frac{(k+q) \cdot x}{2}}$. It would appear as if there would be agreement only for forward scattering. However, it should be recalled that the T -product there was of the form

$T(j_\alpha(\frac{x}{2})j_\beta(-\frac{x}{2}))$. The resolution lies in the fact that there are three possible translations which are natural: (i) by $-y$, (ii) by $-x$, and, (iii) by $\frac{x+y}{2}$. The last yields the representation used by Lehmann, the first yields the representation of Eq. (11.8). The choice (ii) would lead to a representation like Eq. (11.8) but with k replaced by $-q$.

Goldberger also defines another quantity, called M by him. Though he introduces it only for the case of forward scattering, it is possible to introduce it in general as

$$i \int dx e^{-ik \cdot x} \langle p' \lambda' | \{ \theta(x_0) [j_\alpha(x), j_\beta(0)] - \delta(x_0) [j_\alpha(x), \dot{\phi}_\beta(0)] \} | p \lambda \rangle \\ \equiv M_{\alpha\beta}(k, q; p' \lambda', p \lambda) \quad (11.10)$$

Recall that the LSZ formalism gave another representation where the T -products were replaced by the R -products (the retarded commutator). The equivalence of the T -product and R -product representations were exact provided one-particle stability conditions were satisfied. Again, starting from the R -product representation, working out the action of the Klein-Gordon operators but this time on the R -product, and making assumptions similar to what Goldberger made, would indeed reproduce Eq. (11.10). Therefore we should expect the equivalence which Goldberger claimed only for forward scattering to be true for the general case. As we shall see later in this chapter, this was the basis for Salam's proof of dispersion relations for non-forward scattering.

11.4 Forward Scattering Dispersion Relations

11.4.1 Massless Particle Scattering

In this section we address the issue of dispersion relations for the much simpler circumstances of *forward scattering*. We shall see that deriving analyticity from microcausality is far from being straightforward even here. However, the connection is reasonably straightforward in the case of scattering of *massless* particles by massive ones. The classical version of this problem is of course what was tackled by Kramers and Krönig. We shall now show the connection between microcausality and upper half plane analyticity, and consequently the dispersion relations for forward scattering of massless particles follows from field theory. Goldberger makes a further distinction between coherent and incoherent forward scattering; in both cases $k = q$, and hence $p' = p$, but α, β , and, λ, λ' are allowed to be distinct.

For this purpose, let us start by restricting the general considerations discussed above to $\mu = 0$, and further suppressing all internal quantum number indices $\alpha, \beta, \lambda, \lambda'$. More detailed treatments show that the essential connection between microcausality and analyticity in upper half plane of complex frequency is not

compromised by such simplifications. We will start with the analogs of Eq. (11.10) which, for our simplified case, reads as

$$M(k, k; p, p) \equiv i \int dx e^{-ik \cdot x} \langle p | \{ \theta(x_0) [j(x), j(0)] - \delta(x_0) [j(x), \dot{\phi}(0)] \} | p \rangle \quad (11.11)$$

Goldberger claims that the $\delta(x_0)$ term in the above is independent of $k_0 = \omega$. His arguments are based on the fact that equal time commutators are always proportional to $\delta^{(3)}(\mathbf{x})$, which is again tied up with independence of degrees of freedom and microcausality. Rewriting the above equation after dropping the $\delta(x_0)$ terms,

$$M(k, k; p, p) \equiv i \int dx e^{-ik \cdot x} \langle p | \theta(x_0) [j(x), j(0)] | p \rangle \quad (11.12)$$

In many sources, like for example the book by Gasiorowicz [8] (Chap. 22 on analyticity of S-matrix elements), or, in Salam's paper on dispersion relations for non-forward scattering ([9], to be discussed later in this chapter), the above expression entirely in terms of the sources is given as *the* expression for the scattering amplitude, *only* after various assumptions are made about interactions, structure of currents, etc. But as clearly explained above, expressions like Eq. (11.12) follow from the rigorous LSZ-like formalisms, which are non-perturbative, and make very few assumptions about interactions. It is very important to keep this in mind. Now microcausality demands $[j(x), j(0)]$ should vanish for *space-like* x_μ . That is, for $x^2 < 0$. This, as well as the retardation condition $\theta(x_0)$ in the expression for the amplitude will prove to be crucial. This means that the forward scattering amplitude in this case can be represented as (Goldberger's metric convention is $(-+++)$)

$$\int_0^\infty dx_0 \int d\mathbf{x} e^{i\omega(x_0 - \hat{\mathbf{k}} \cdot \mathbf{x})} f(x_0, \mathbf{x}) \quad (11.13)$$

with

$$f(x_0, \mathbf{x}) = 0 \quad x_0 < |\mathbf{x}| \quad (11.14)$$

As it stands, $f(x_0, \mathbf{x})$ has still a dependence on p ; following Goldberger, let us choose to work in the frame in which $\mathbf{p} = 0$. Then the amplitude depends only on the “photon energy” ω . Since the amplitudes have to only depend on Lorentz invariants, there is no loss of generality in this. If additional aspects like spin had been taken into account, this discussion would get more involved, but in the end it is only invariant amplitudes and their dependences on invariants that matter.

Now it is clear that the integrations in Eq. (11.14) are over regions with $x_0 - \hat{\mathbf{k}} \cdot \mathbf{x} > 0$. This immediately yields the crucial result that if the frequency ω is extended to the *complex* plane, the amplitude is still well defined for $\text{Im } \omega > 0$. In other words, the representation above defines an analytic continuation $F(\omega)$ of the amplitude to the upper half of the complex- ω plane (see also [8] Chap. 22 for a lucid discussion).

We can just repeat the steps employed in the earlier discussion of the Kramers-Krönig dispersion relations. The Cauchy residue theorem for contour integrations gives

$$\operatorname{Re} F(\omega) = \frac{1}{\pi} P \int_{-\infty}^{\infty} d\omega' \frac{\operatorname{Im} F(\omega')}{\omega' - \omega} \quad (11.15)$$

This raises the question of the convergence of the integral on the r.h.s. This issue has already cropped up before and is not something that analyticity on its own can resolve. It depends on the dynamical details about the system in question. In case the integral does not converge, one has to resort to the so-called subtractions. The number of such subtractions is determined by the severity of the large ω dependence of $\operatorname{Im} F(\omega)$. We shall return to this when we discuss applications of dispersion relations to problems of actual physical interest.

Another important issue is of expressing the r.h.s in terms of measurable quantities. While the imaginary part of the amplitude is proportional to the total cross section which is an observable, it is so only for positive ω , which can be identified as the physical region. But the integral on the r.h.s extends over both positive and negative values of ω . However, it is easy to verify directly from the expression for the amplitude that

$$F^*(\omega) = F(-\omega) \rightarrow \operatorname{Im} F(\omega) = -\operatorname{Im} F(-\omega) \quad (11.16)$$

In the larger context, this is the analog of *Crossing Symmetry*, which we saw emerged naturally from the LSZ formalism. Using this, the dispersion relation can be recast as

$$\operatorname{Re} F(\omega) = \frac{2}{\pi} P \int_0^{\infty} d\omega' \frac{\omega' \operatorname{Im} F(\omega')}{\omega'^2 - \omega^2} \quad (11.17)$$

11.4.2 Massive Particle Scattering: Goldberger Analysis

The moment we try to generalize these considerations to massive particle scatterings, even the forward case becomes quite subtle and involved. The major difficulty is that the exponent is now of the form $\omega x_0 - \sqrt{\omega^2 - \mu^2} \hat{\mathbf{k}} \cdot \mathbf{x}$ and this is not *uniform* in ω . This innocuous looking difference causes all the technical complications. Capps and Takeda [10] proceed by assuming uniform boundedness in the upper half plane and show dispersion relations follow. This approach is too mathematical and the basis for analyticity as well as the domains of analyticity are rather obscure. We shall instead follow Goldberger's treatment [6] as it is more transparent, and, as we shall see in a later section, amenable to a generalization to the more general, and even more difficult, problem of dispersion relations for non-forward scattering (Salam's work).

The starting point for Goldberger's treatment is Eq. (11.10). As stated before, the $M_{\alpha\beta}$ is completely equivalent to the amplitude one obtains from the LSZ formalism, also called the Feynman matrix element by Goldberger, for *positive* values of $k_0 = \omega$. There are a number of intermediate steps used by Goldberger in arriving at the final

result. They are mostly elementary to derive, but very important for the derivation. The three relations used by him are (his Eq. (2.11)):

$$\begin{aligned} M_{\alpha\beta}(\mathbf{k}, \omega; \lambda', \lambda) &= M_{\alpha\beta}(-\mathbf{k}, \omega; \lambda', \lambda) \\ \langle p\lambda' | [j_\alpha(-x), j_\beta(0)] | p\lambda \rangle &= \langle p\lambda' | [j_\alpha(0), j_\beta(x)] | p\lambda \rangle \\ M_{\alpha\beta}(-\omega; \lambda', \lambda) &= M_{\alpha\beta}(\omega; \lambda, \lambda')^* \end{aligned} \quad (11.18)$$

The first of these follows from *Parity Invariance*, expected to be good for strong interactions (in any case, at the time of Goldberger's paper, April 1955, even the Parity Violation in nuclear β -decay had not been discovered. Wu's experiment demonstrating maximal parity violation was performed in December 1956). Goldberger chooses to work in the *laboratory frame* where the nucleon momentum $\mathbf{p} = 0$. Then the first condition implies that $M_{\alpha\beta}$ can only depend on powers of $\mathbf{k}^2 = \omega^2 - \mu^2$, and it is a function of only ω (for every $\alpha, \beta, \lambda, \lambda'$). The last of the three equations reflects this. The second of these equations follows directly from *translational invariance*.

An important part of Goldberger's proof is to split both $M_{\alpha\beta}$, $F_{\alpha\beta}$ into their *dispersive* and *absorptive* parts. The physical basis for this decomposition is whether *real* intermediate states are involved (dispersive), or, *virtual* intermediate states are involved (absorptive). This technique has been used by many in the study of dispersion relations. Writing these explicitly:

$$\begin{aligned} F_{\alpha\beta}(k; \lambda', \lambda) &= D_{\alpha\beta}(k; \lambda', \lambda) + i\epsilon(k_0)A_{\alpha\beta}(k; \lambda', \lambda) \\ M_{\alpha\beta}(k; \lambda', \lambda) &= D_{\alpha\beta}(k; \lambda', \lambda) + iA_{\alpha\beta}(k; \lambda', \lambda) \end{aligned} \quad (11.19)$$

Note the important difference between $M_{\alpha\beta}$ and $F_{\alpha\beta}$ involving the factor $\epsilon(k_0)$. Therefore, only for $k_0 > 0$ do these two agree with each other. As explicitly remarked by Goldberger, the third of Eq. (11.18) is only satisfied by $M_{\alpha\beta}$ and not $F_{\alpha\beta}$ due to the above-mentioned differences. Another consequence of the first of Eq. (11.18) is that $e^{-i\mathbf{k}\cdot\mathbf{x}}$ in Eq. (11.10) can be replaced by $\cos \mathbf{k} \cdot \mathbf{x}$, leading to

$$\begin{aligned} D_{\alpha\beta}(\omega; \lambda', \lambda) &= \frac{i}{2} \int dx e^{i\omega x_0} \cos \mathbf{k} \cdot \mathbf{x} \langle p\lambda' | X_{\alpha\beta}(x) | p\lambda \rangle \\ X_{\alpha\beta}(x) &= \{\epsilon(x_0)[j_\alpha(x), j_\beta(0)] - 2\delta(x_0)[j_\alpha(x), \dot{\phi}_\beta(0)]\} \\ A_{\alpha\beta}(\omega; \lambda', \lambda) &= \frac{1}{2} \int dx e^{i\omega x_0} \cos \mathbf{k} \cdot \mathbf{x} \langle p\lambda' | [j_\alpha(x), j_\beta(0)] | p\lambda \rangle \end{aligned} \quad (11.20)$$

It is very important to note that the essential difference between these expressions (apart from the equal time term proportional to $\delta(x_0)$ which will be seen to be irrelevant in the context of dispersion relations) is the factor $\epsilon(x)$ (the dispersive part has the factor); it is amazing that the entire distinction rests on this simple factor! Goldberger then introduces

$$\begin{aligned} 2K_{\alpha\beta}^{(+)}(x) &= [j_\alpha(x), j_\beta(0)] + [j_\alpha(0), j_\beta(x)] \\ 2K_{\alpha\beta}^{(-)}(x) &= [j_\alpha(x), j_\beta(0)] - [j_\alpha(0), j_\beta(x)] \end{aligned} \quad (11.21)$$

so that

$$[j_\alpha(x), j_\beta(0)] = K_{\alpha\beta}^{(-)}(x) + K_{\alpha\beta}^{(+)}(x) \quad (11.22)$$

Viewed as matrix elements or in other words using the shorthand notation $K_{\alpha\beta}^{(\pm)}(x)$ for $\langle p\lambda' | K_{\alpha\beta}^{(\pm)}(x) | p\lambda \rangle$, the following symmetry properties follow directly from the second of Eq. (11.18):

$$K_{\alpha\beta}^{(-)}(x) = K_{\beta\alpha}^{(-)}(x) \quad K_{\alpha\beta}^{(+)}(x) = -K_{\beta\alpha}^{(+)}(x) \quad (11.23)$$

Additionally

$$K_{\alpha\beta}^{(+)}(x) = K_{\alpha\beta}^{(+)}(-x) \quad K_{\alpha\beta}^{(-)}(x) = -K_{\alpha\beta}^{(-)}(-x) \quad (11.24)$$

It is obvious that these symmetry properties cannot hold at an operator level. What Goldberger considers is a case where the scattering amplitude (F , equivalently M) is a *matrix* labelled by α, β . Therefore one expects a *matrix form* of the dispersion relations. Hence it makes sense to separately consider the *symmetric* and *antisymmetric* parts. The symmetry properties above facilitate such a split. The next step in the algebra is therefore to further split the $D_{\alpha\beta}, A_{\alpha\beta}$ into their symmetric and antisymmetric parts:

$$\begin{aligned} D_{\alpha\beta}(\omega; \lambda', \lambda) &= D_{\alpha\beta}^{(1)}(\omega; \lambda', \lambda) + i D_{\alpha\beta}^{(2)}(\omega; \lambda', \lambda) \\ A_{\alpha\beta}(\omega; \lambda', \lambda) &= A_{\alpha\beta}^{(1)}(\omega; \lambda', \lambda) + i A_{\alpha\beta}^{(2)}(\omega; \lambda', \lambda) \end{aligned} \quad (11.25)$$

Here, the superscripts 1, 2 label the symmetric and antisymmetric (in $\alpha\beta$) parts respectively. The explicit expressions for these four quantities are easily found:

$$\begin{aligned} D_{\alpha\beta}^{(1)} &= \frac{i}{2} \int dx \cos \mathbf{k} \cdot \mathbf{x} \cos \omega x_0 \langle p\lambda' | \\ &\quad \cdot \{ \epsilon(x_0) K_{\alpha\beta}^{(-)}(x) - 2\delta(x_0) [j_\alpha(x), \dot{\phi}_\beta(0)] \} | p\lambda \rangle \\ D_{\alpha\beta}^{(2)} &= \frac{i}{2} \int dx \cos \mathbf{k} \cdot \mathbf{x} \sin \omega x_0 \langle p\lambda' | \epsilon(x_0) K_{\alpha\beta}^{(+)}(x) | p\lambda \rangle \end{aligned} \quad (11.26)$$

and,

$$\begin{aligned} A_{\alpha\beta}^{(1)} &= \frac{i}{2} \int dx \cos \mathbf{k} \cdot \mathbf{x} \sin \omega x_0 \langle p\lambda' | K_{\alpha\beta}^{(-)}(x) | p\lambda \rangle \\ A_{\alpha\beta}^{(2)} &= -\frac{i}{2} \int dx \cos \mathbf{k} \cdot \mathbf{x} \cos \omega x_0 \langle p\lambda' | K_{\alpha\beta}^{(+)}(x) | p\lambda \rangle \end{aligned} \quad (11.27)$$

An aspect of these expressions that will play a crucial role subsequently needs some elaboration. This is the correlation between symmetry (in $\alpha\beta$) and the trigonometric functions. For example, $A_{\alpha\beta}^{(1)}$ which is symmetric, has $\sin \omega x_0$ in its integrand,

while the antisymmetric part $A_{\alpha\beta}^{(2)}$ has a $\cos \omega x_0$ in its integrand. This correlation gets reversed for $D_{\alpha\beta}$ where the symmetric $D_{\alpha\beta}^{(1)}$ comes with $\cos \omega x_0$ while the antisymmetric $D_{\alpha\beta}^{(2)}$ comes with $\sin \omega x_0$! The appearance of trigonometric functions in place of the original $e^{i\omega x_0}$ may have given the erroneous impression that these are the real and imaginary parts.

Due to the extreme importance of this correlation, we explicitly show how this comes about. Let us consider $A_{\alpha\beta}^{(1)}$ which is symmetric. Consequently it has $\langle p\lambda' | K_{\alpha\beta}^{(-)}(x) | p\lambda \rangle$. Let us trivially rewrite this as (we have used the earlier mentioned shorthand notation for matrix elements)

$$A_{\alpha\beta}^{(1)} = \frac{1}{2} \cdot \frac{1}{2} \int d^4x e^{i\omega x_0} \cos \mathbf{k} \cdot \mathbf{x} \{K_{\alpha\beta}^{(-)}(x) + K_{\alpha\beta}^{(-)}(x)\} \quad (11.28)$$

On using $K_{\alpha\beta}^{(-)}(x) = -K_{\alpha\beta}^{(-)}(-x)$ in the second part, changing variable of integration to $-x$, it is easy to see that this becomes

$$A_{\alpha\beta}^{(1)} = \frac{1}{2} \cdot \frac{1}{2} \int d^4x (e^{i\omega x_0} - e^{-i\omega x_0}) \cos \mathbf{k} \cdot \mathbf{x} K_{\alpha\beta}^{(-)}(x) \quad (11.29)$$

and we see the emergence of $-i \sin \omega x_0$! Similar steps lead to the occurrence of $\cos \omega x_0$ in $A_{\alpha\beta}^{(2)}$. The reversal of this correlation between symmetry and trigonometric functions for $D_{\alpha\beta}^{(1,2)}$ is easily understood as for them there is the additional factor of $\epsilon(x)$ which is an *odd* function, i.e. $\epsilon(x) = -\epsilon(-x)$. It is somewhat of an irony that in order to arrive at the Dispersion Relations Goldberger has to go back to the $e^{i\omega x_0}$ form! More on this later.

The trigonometric functions also make manifest the oddness (evenness) of the four quantities as functions of ω which too will prove critical in arriving at the Dispersion Relations.

$$\begin{aligned} A_{\alpha\beta}^{(1)}(\omega) &= -A_{\alpha\beta}^{(1)}(-\omega); A_{\alpha\beta}^{(2)}(\omega) = A_{\alpha\beta}^{(2)}(-\omega) \\ D_{\alpha\beta}^{(1)}(\omega) &= D_{\alpha\beta}^{(1)}(-\omega); D_{\alpha\beta}^{(2)}(\omega) = -D_{\alpha\beta}^{(2)}(-\omega) \end{aligned} \quad (11.30)$$

We state here the three main ingredients that led to these results: (i) the space-time dimensionality is even (in our case $d = 4$) so that under $x \rightarrow -x$ the measure d^4x remained unchanged, (ii) the $\cos \mathbf{k} \cdot \mathbf{x}$ factor which is insensitive to the parity part, $\mathbf{x} \rightarrow -\mathbf{x}$, of the $x \rightarrow -x$. This arose from the assumed parity invariance of scattering amplitudes, and finally, (iii) the oddness and evenness of the $K_{\alpha\beta}^{(\pm)}(x)$ under $x \rightarrow -x$.

It is important to note that the x_0 -integrations are restricted to positive values of x_0 only. Goldberger shows that in the absence of *discrete bound states*, $A_{\alpha\beta}(\omega; \lambda', \lambda)$ *vanishes* for $|\omega| < \mu$. Furthermore, he also shows that there are branch points for $|\omega| > \mu$ corresponding to thresholds for real processes. Goldberger shows this by

the usual trick of introducing a complete set of states of the interacting system and then carrying out the x_0 -integration.

Another important remark, which has already been invoked, is that the terms involving $\delta(x_0)$ are *independent* of ω , as equal time commutators are always proportional to $\delta^{(3)}(\mathbf{x})$.

Goldberger then considers the integral

$$I_{\alpha\beta}(\omega, \omega_0; \lambda', \lambda) \equiv \frac{\omega^2 - \omega_0^2}{\pi} \int_{-\infty}^{\infty} d\omega' \frac{A_{\alpha\beta}^{(1)}(\omega'; \lambda', \lambda) + i A_{\alpha\beta}^{(2)}(\omega'; \lambda', \lambda)}{(\omega' - \omega_0)(\omega'^2 - \omega^2)} \quad (11.31)$$

The reader may wonder about the significance of the additional variable ω_0 , and the need for its introduction. This will become clear towards the end of this discussion. The integral expressions for the $A_{\alpha\beta}$ are now substituted into this, and the order of space-time and ω' integrations are interchanged. Details and justifications for the interchanging of integrations can be found in Goldberger's paper. I provide here a few more intermediate steps to facilitate the reader. On using the oddness-evenness of $A_{\alpha\beta}$ as functions of ω , this integral can be rewritten as

$$I_{\alpha\beta}(\omega, \omega_0; \lambda', \lambda) \equiv \frac{\omega^2 - \omega_0^2}{\pi} \int_{-\infty}^{\infty} d\omega' \frac{\omega' A_{\alpha\beta}^{(1)}(\omega'; \lambda', \lambda) + i\omega_0 A_{\alpha\beta}^{(2)}(\omega'; \lambda', \lambda)}{(\omega'^2 - \omega_0^2)(\omega'^2 - \omega^2)} \quad (11.32)$$

After the said interchange of the order of integrations, it is easy to see that

$$I_{\alpha\beta}(\omega, \omega_0; \lambda', \lambda) = i \int_0^{\infty} dx_0 \int d\mathbf{x} \langle p\lambda' | X_{\alpha\beta} | p\lambda \rangle$$

$$X_{\alpha\beta} = \{K_{\alpha\beta}^{(-)}(x)J^{(1)}(x; \omega, \omega_0) + i K_{\alpha\beta}^{(+)}(x)J^{(2)}(x; \omega, \omega_0)\} \quad (11.33)$$

Here $J^{(1,2)}$ are given by

$$J^{(1)}(x; \omega, \omega_0) = \frac{\omega^2 - \omega_0^2}{\pi} \int_{-\infty}^{\infty} d\omega' \frac{\omega' \sin \omega' x_0 \cos \mathbf{k}' \cdot \mathbf{x}}{(\omega'^2 - \omega_0^2)(\omega'^2 - \omega^2)} \quad (11.34)$$

and,

$$J^{(2)}(x; \omega, \omega_0) = -\frac{\omega^2 - \omega_0^2}{\pi} \int_{-\infty}^{\infty} d\omega' \frac{\omega_0 \cos \omega' x_0 \cos \mathbf{k}' \cdot \mathbf{x}}{(\omega'^2 - \omega_0^2)(\omega'^2 - \omega^2)} \quad (11.35)$$

Equation (2.28) of Goldberger's paper has ω_0 factor in the numerator missing. So far, no approximations have been made except for the assumption about interchanging the order of integrations. Now Goldberger proceeds to evaluate the integrals $J^{(1,2)}$. These look like the kinds that can be solved routinely through contour integration techniques. The integrands have simple poles at $\pm \omega_0, \pm \omega$. Here comes the subtlety about the J-integrals. Goldberger remarks that in the integration region of interest,

i.e. $x_0 > |\mathbf{x}|$, $\sin \omega x_0$ and $\cos \omega x_0$ have to be replaced by $-i e^{i\omega x_0}$ and $e^{i\omega x_0}$ respectively (no, there is no mistake of a factor of 2!). This is a crucial step in Goldberger's derivation, but unfortunately his rationalization of this step is not very convincing. In fact the other exponential in these trigonometric functions is $e^{-i\omega x_0}$ and it actually diverges in the upper half ω -plane for $x_0 > 0$! With both of the trigonometric functions in the integrand, the contours cannot be closed in either the upper or the lower half-planes.

We offer a different explanation for these replacements. In the case of $J^{(1)}$, which involves $\sin \omega' x_0$, what multiplies $\sin \omega' x_0$ in the integrand is an *odd* function of ω' . Therefore, if we write $\sin \omega' x_0 = -\frac{i}{2}(e^{i\omega' x_0} - e^{-i\omega' x_0})$, in the part with $e^{-i\omega' x_0}$, one can change variable of integration from ω' to $-\omega'$ to get an identical contribution as the part with $e^{i\omega' x_0}$, thus justifying the replacement $\sin \omega' x_0 \rightarrow -i e^{i\omega' x_0}$. In the case of the $J^{(2)}$, the integrand has $\cos \omega' x_0$ multiplied by an *even* function of ω' , and a similar reasoning leads to the replacement of $\cos \omega' x_0 \rightarrow e^{i\omega' x_0}$. The reader should go back to where we showed how the trigonometric functions arose. That was just a reversal of what we did just now and both depended on the same three ingredients listed earlier.

With these crucial replacements, the J 's can be evaluated by methods of contour integration in the upper half plane. The integrands for both of them have simple poles on the real axis at $\omega' = \pm \omega_0, \pm \omega$. Results given by Goldberger are

$$\begin{aligned} J^{(1)}(x; \omega, \omega_0) &= \cos \omega x_0 \cos \mathbf{k} \cdot \mathbf{x} - \cos \omega_0 x_0 \cos \mathbf{k}_0 \cdot \mathbf{x} \\ J^{(2)}(x; \omega, \omega_0) &= \omega_0 \left\{ \frac{1}{\omega} \sin \omega x_0 \cos \mathbf{k} \cdot \mathbf{x} - \frac{1}{\omega_0} \sin \omega_0 x_0 \cos \mathbf{k}_0 \cdot \mathbf{x} \right\} \quad (11.36) \end{aligned}$$

We have rewritten the expression for $J^{(2)}$ so as to bring out the symmetry between ω and ω_0 explicitly. Some important remarks are in order at this stage. The criticality of replacing the trigonometric functions by a single exponential becomes evident at this point. It is only because of that step that $J^{(1)}$ whose integrand had the factor $\sin \omega' x_0$ becomes of the form $X(\omega) - X(\omega_0)$ where $X(\omega)$ itself is proportional to $\cos \omega x_0$. Likewise, $J^{(2)}$ whose integrand was proportional to $\cos \omega' x_0$ becomes of the form $Y(\omega) - Y(\omega_0)$ where $Y(\omega)$ is proportional to $\sin \omega x_0$. In order to appreciate the significance of this, note that the integrand of $A_{\alpha\beta}^{(1)}$ was proportional to $\sin \omega x_0 K_{\alpha\beta}^{(-)}(x)$, and $A_{\alpha\beta}^{(2)}$ to $\cos \omega x_0 K_{\alpha\beta}^{(+)}(x)$; in contrast, the dispersive part $D_{\alpha\beta}^{(1)}$ had (apart from the $\delta(x_0)$ piece which is *independent* of ω) an integrand proportional to $\cos \omega x_0 K_{\alpha\beta}^{(-)}(x)$ while $D_{\alpha\beta}^{(2)}$ had an integrand proportional to $\sin \omega x_0 K_{\alpha\beta}^{(+)}(x)$. Thus the replacement of the trigonometric functions by the $e^{i\omega x_0}$, which interchanged the trigonometric functions upon integration, has the net effect of relating certain integrals of the absorptive parts $A_{\alpha\beta}^{(1,2)}$ directly to the dispersive parts $D_{\alpha\beta}^{(1,2)}$. But that is precisely the content of the dispersion relations.

Putting everything together, the integral $I_{\alpha\beta}$ is found to be

$$I_{\alpha\beta}(\omega, \omega_0; \lambda', \lambda) = D_{\alpha\beta}^{(1)}(\omega; \lambda', \lambda) - D_{\alpha\beta}^{(1)}(\omega_0; \lambda', \lambda) + i\omega_0 \left\{ \frac{D_{\alpha\beta}^{(2)}(\omega; \lambda', \lambda)}{\omega} - \frac{D_{\alpha\beta}^{(2)}(\omega_0; \lambda', \lambda)}{\omega_0} \right\} \quad (11.37)$$

This is the full and explicit statement of what was expressed verbally earlier; integrals over $A_{\alpha\beta}^{(1,2)}$ being expressed as (difference of) $D_{\alpha\beta}^{(1,2)}$. It is important to appreciate that the ω -independence of the terms in $D_{\alpha\beta}^{(1)}$ proportional to $\delta(x_0)$ is crucial in arriving at this.

Separating the above (we have used Eq. (11.32)) into its symmetric and antisymmetric parts (in $\alpha\beta$):

$$D_{\alpha\beta}^{(1)}(\omega; \lambda', \lambda) - D_{\alpha\beta}^{(1)}(\omega_0; \lambda', \lambda) = \frac{2(\omega^2 - \omega_0^2)}{\pi} \int_0^\infty d\omega' \omega' \frac{A_{\alpha\beta}^{(1)}(\omega'; \lambda', \lambda)}{(\omega'^2 - \omega_0^2)(\omega'^2 - \omega^2)} \quad (11.38)$$

$$\frac{D_{\alpha\beta}^{(2)}(\omega; \lambda', \lambda)}{\omega} - \frac{D_{\alpha\beta}^{(2)}(\omega_0; \lambda', \lambda)}{\omega_0} = \frac{2(\omega^2 - \omega_0^2)}{\pi} \int_0^\infty d\omega' \omega' \frac{\frac{A_{\alpha\beta}^{(2)}(\omega'; \lambda', \lambda)}{\omega'}}{(\omega'^2 - \omega_0^2)(\omega'^2 - \omega^2)} \quad (11.39)$$

We recognize these as the *once subtracted* form of dispersion relations, separately for the symmetric and antisymmetric parts. The subtraction is done at the point $\omega = \omega_0$. This explains the significance of ω_0 introduced in Eq. (11.31). The limits of ω' -integrations have been brought to the physical region $(0, \infty)$ by exploiting the oddness (evenness) of $A_{\alpha\beta}^{(1,2)}$ as functions of ω .

It is tempting to write down the *unsubtracted* form of these (Goldberger does not do it) provided the high frequency behaviours assure their existence:

$$D_{\alpha\beta}^{(1)}(\omega; \lambda', \lambda) = \frac{2}{\pi} \int_0^\infty d\omega' \omega' \frac{A_{\alpha\beta}^{(1)}(\omega'; \lambda', \lambda)}{(\omega'^2 - \omega^2)} \quad (11.40)$$

and,

$$\frac{D_{\alpha\beta}^{(2)}(\omega; \lambda', \lambda)}{\omega} = \frac{2}{\pi} \int_0^\infty d\omega' \omega' \frac{\frac{A_{\alpha\beta}^{(2)}(\omega'; \lambda', \lambda)}{\omega'}}{(\omega'^2 - \omega^2)} \quad (11.41)$$

That these two forms of dispersion relations are not identical and should not cause any problems; as already stated, the symmetric and antisymmetric can have independent structures. Nevertheless, the second of these can, in principle, have the same structure as the first.

The concluding parts of Goldberger's paper deal with the important issue of relating the absorptive part of the amplitude to directly observable (measurable)

quantities. This is an important part of all dispersion relations based approaches. We refer the reader to Goldberger's paper for details.

As we have seen, Goldberger does not explicitly address questions of domains of analyticity. He very cleverly arrives at the Hilbert transforms directly. If Toll's ideas of logical equivalence should hold in the RQFT context too, then the dispersion relations would be logically equivalent to the analyticity aspects. In that sense, Eq. (11.41) would be logically equivalent to $\frac{D_{\alpha\beta}^{(2)} + i A_{\alpha\beta}^{(2)}}{\omega}$ being an analytical function in the upper half plane and it is compatible with $D_{\alpha\beta}^{(2)} + i A_{\alpha\beta}^{(2)}$ itself being an analytical function in the upper half plane, exactly akin to the logical equivalence of Eq. (11.40) with a similar analyticity of $D_{\alpha\beta}^{(1)} + i A_{\alpha\beta}^{(1)}$.

11.4.3 Massive Particle Scattering: Symanzik Analysis

Now we discuss Kurt Symanzik's treatment of Dispersion Relations for forward scattering of *massive* particles [11]. This is the same problem whose treatment by Goldberger we have discussed at length, but there are many differences in the two approaches. Firstly, Symanzik's metric convention is $(+ - - -)$, same as that of Barton, Gasirowicz and Bjorken and Drell, but opposite that of Goldberger.

Symanzik treats the scattering of two neutral scalars of masses M (called nucleons) and μ (called mesons) with $M > \mu$. The nucleon field is $\psi(x)$ while the meson field is $\phi(x)$. Symanzik imposes a *parity* selection rule on the nucleons: $\langle \text{even} | \psi | \text{even} \rangle = \langle \text{odd} | \psi | \text{odd} \rangle = 0$. Such selection rules, whether for nucleons or mesons or both are primarily intended for appropriately restricting masses of multiparticle threshold states. The analyticity domains are sensitive to masses. Though such selection rules restrict the generality of the analyses, they do so in rather mild ways.

For the scattering $Nucleon(p) + Meson(k) \rightarrow Nucleon(p') + Meson(k')$, the S-matrix element is denoted by Symanzik as $\langle p' k' | S | p, k \rangle$. He follows the LSZ reduction techniques, but with some novel twists. He chooses to reduce in the nucleons unlike Goldberger above who chose to reduce in the mesons. For the case in question the scattering amplitude is not a matrix in internal space, and the many novel features of Goldberger's analysis on this count do not show up here. The nucleon source is denoted by $O(x)$, i.e. $O(x) = \mathcal{K}_{x,M} \psi(x)$. After the first reduction of the nucleon in final state with p' , he obtains

$$\langle p' k' | S | p k \rangle = \langle p' k' | p k \rangle + i(2\pi)^4 \delta^{(4)}(p' + k' - p - k) \langle k' | O(0) | p k \rangle \quad (11.42)$$

If we recall our earlier discussions of LSZ reductions, this looks different, with no space-time integrations while the LSZ reduction at this stage had one space-time integration. It is worth going into some details here. The LSZ formalism would have given, for this reduction:

$$\langle p' k' | S | p k \rangle = \langle p' k' | p k \rangle + i \int dy e^{ip' \cdot y} \langle k' | O(y) | p k \rangle \quad (11.43)$$

Using the identity for translations

$$O(y) = e^{iP \cdot y} O(0) e^{-P \cdot y} \quad (11.44)$$

the non-trivial part of the S-matrix element (denoted with a ι) becomes

$$\langle p', k' | S | pk \rangle' = i \int dy e^{ip' \cdot y} e^{ik' \cdot y} e^{-i(p+k) \cdot y} \langle k' | O(0) | pk \rangle \quad (11.45)$$

Performing the y -integration, one gets

$$\langle p' k' | S | pk \rangle' = i(2\pi)^4 \delta^{(4)}(p' + k' - p - k) \langle k' | O(0) | pk \rangle \quad (11.46)$$

So, Symanzik's Eq. (11.42) is indeed equivalent to the results from LSZ reductions. As a matter of fact, Symanzik's form is much better as it does away with one space-time integration. Subsequent reductions starting from this will have one less space-time integration too, without having to go through the steps we went through to achieve that. Also, Symanzik's form allows one to peel off the expected energy-momentum conservation delta function in the first step itself, leading to the T -matrix element:

$$T(p' k'; pk) = \langle k' | O(0) | pk \rangle \quad (11.47)$$

From here, the subsequent reductions proceed exactly as in the LSZ formalism (what else). Symanzik again reduces the nucleon in the initial state with momentum p . The asymptotic pieces, i.e. $x_0 \rightarrow \pm\infty$ again vanish, though for different reasons. In the pure-meson case, the initial and final states are two identical particle states, while in the meson-nucleon scattering case both of them have only one particle of each kind. The one-particle stability conditions as also the action of annihilation operators work differently. The reader is encouraged to work out these details.

As explained before, one can either use the T -product representation, or the R -product representation. Either way, one can, after working out the action of the Klein-Gordon operator on $\theta(x_0)$ as well as the fields, to end up with either T or R -products of sources plus terms with $\delta(x_0)$ (equal time) and higher derivatives of the delta functions. In Goldberger's analysis, such terms did not play any role in the derivation of the Dispersion Relations. Symanzik argues that such terms are necessarily of the form $\mathcal{R}((p \cdot k)^2)$ which is a *polynomial* in its arguments and hence *real*. Modulo this term, Symanzik arrives at the expression for the T -matrix element, now specialized to the *forward scattering* (his Eq. (6)):

$$T(pk; pk) = i \int dx e^{ip \cdot x} \theta(x_0) \langle k | [O(x), O(0)] | k \rangle \quad (11.48)$$

Now Symanzik chooses to work in the rest frame of the meson in which $\omega = \frac{p \cdot k}{\mu}$ is the energy of the nucleon. He carries out the *angular* integrations first to obtain (the T -matrix element is only a function of ω now)

$$T(\omega)' = \int_0^\infty F(\omega, r) dr \quad (11.49)$$

where

$$F(\omega, r) = 4\pi i r (\omega^2 - M^2)^{-\frac{1}{2}} \sin \sqrt{\omega^2 - M^2} r \cdot \int_0^\infty dt e^{i\omega t} \langle \mu, 0 | [O(\mathbf{x}, t), O(0)] | \mu, 0 \rangle \quad (11.50)$$

At this point Symanzik claims that $F(\omega, r)$ given above is definable for all values of ω on the real axis (not just for $\omega^2 > M^2$ as one might naively conclude). This is an extremely important claim as it fixes the domain of analyticity for $T(\omega)$, which is central to the derivation of the Dispersion Relations. But unfortunately his reasoning is not very transparent at this critical juncture. His point seems to be (I urge the reader to independently assess the matter) that one can rewrite $F(\omega, r)$ as

$$F(\omega, r) = 4\pi i r (\omega^2 - M^2)^{-\frac{1}{2}} \sin \sqrt{\omega^2 - M^2} r e^{i\omega r} \cdot \int_0^\infty dt e^{i\omega(t-r)} \langle \mu, 0 | [O(\mathbf{x}, t), O(0)] | \mu, 0 \rangle \quad (11.51)$$

The rescaled integral is regular in the upper half complex ω -plane as long as $t - r > 0$ which is guaranteed by the microcausality condition. What happens exactly on the light cone $t = r$ is a subtler issue. Symanzik states that if one makes the reasonable assumption that $\langle \mu, 0 | [O(\mathbf{x}, t), O(0)] | \mu, 0 \rangle$ is *continuous* inside the light cone, the asymptotic behaviour of $F(\omega, r)$ as $\omega \rightarrow \infty$ is controlled by the singularities of the said matrix element of the commutator *on* the light cone. He ascribes this result to the Riemann-Lebesgue Lemma.

Then what remains depends on the regularity of the remaining factors in the upper half plane, possibly including the real axis. Symanzik claims that they are regular. His reasoning being that the exponentially increasing part of $\sin \sqrt{\omega^2 - M^2} r$ is “just compensated” by the exponentially decreasing part of $e^{i\omega r}$. The fact that r is always positive being essential to this reasoning. But this compensation can only happen provided

$$\text{Im } \omega > |\text{Im } \sqrt{\omega^2 - M^2}| \quad (11.52)$$

From this, I don’t see any straightforward way of showing that the domain of analyticity is the upper half plane for (complex) ω that includes the real axis. We can give the simple counter example of points in this domain that violate Eq. (11.52). Take all the points lying on the imaginary axis, i.e. $\omega = i\omega_I$. Then the above condition translates to $\omega_I > |\sqrt{\omega_I^2 + M^2}|$, which is obviously impossible to fulfil.

In fact, unless the upper half plane domain of analyticity can be systematically shown to emerge from the above constraint, Symanzik’s proof should be considered *incomplete*. Later we shall see almost identical issues arising when we discuss dispersion relations for non-forward scattering. Goldberger’s proof does not encounter these difficulties.

Despite this apparent stumbling block, we can continue with Symanzik’s line of reasoning. Of course, once the domain of analyticity is accepted to be the upper half

plane, it may seem trivial to write down the dispersion relations as straightforward consequences of the Cauchy Residue Theorem. But there is a catch here. Even after accepting Symanzik's arguments, one only has upper half plane analyticity for $F(\omega, r)$ but not necessarily for $T'(\omega)$ as the latter is an integral of the former over r . The point is that while for *real* ω this integral exists, it is not at all obvious it does so for every *complex* ω in the upper half plane. Symanzik goes through some pretty elaborate manipulations to settle this issue. We refer the interested reader to Symanzik's paper for details, but simply quote his final conclusions.

He says his Eq. (15) is just the Goldberger dispersion relation for his model. In particular, that the scattering amplitude T' as a function of $\omega' = \frac{\mu}{M} \cdot \omega$, which is the meson energy in the lab frame defined by nucleons at rest, may be analytically continued from $\omega' \geq \mu$ to the entire cut plane with cuts from $-\infty$ to $-\mu$, and, μ to ∞ . It has poles at $\omega' = \pm \frac{\mu^2}{2M}$. The analytically continued amplitude satisfies $T'(\omega') = (T'(-\omega'^*))^*$. This is nothing but one of the crossing symmetry relations.

There are several interesting features of Symanzik's analysis; we merely state two of them, and refer the reader to the paper for more details. First of these is his remark that reducing the heavier field in the LSZ formalism leads to certain essential simplifications; the second is that Symanzik extends his analysis to derive dispersion relations also for derivatives of the scattering amplitude wrt scattering angle evaluated at zero scattering angle. From an experimental point of view exact forward scattering is rather exceptional. So information about as many derivatives of the scattering amplitude wrt scattering angle at zero scattering angle gives more and more information and scattering at very small angles can be feasible. This is a step towards dispersion relations at finite scattering angles, a topic we will discuss in the next section.

11.5 Non-forward Scattering: Salam's Approach

The next level of generality in dispersion relations would be to consider *non-forward scattering* of massive particles. From Goldberger's thorough analysis (as also Symanzik's) it is quite clear that relaxing the restriction of forward scattering is going to be rather formidable! For generic scattering angles (we shall continue to look at the simplest case of $2 \rightarrow 2$ scatterings), say, $N(p) + M(k) \rightarrow N(p') + M(k')$ discussed earlier, even after neglecting spin complications, there are two *independent* invariants that are needed to describe the scattering process. Called *Mandelstam Variables*, these are $s = (p + k)^2$, $t = (p - p')^2$, $u = (p - k')^2$ subject to the overall energy-momentum conservation $p + k = p' + k'$. Because of that, s, t, u are not independent; without requiring the particles to be *on shell*, i.e. satisfying $p_i^2 = m_i^2$, they obey the constraint, in this particular case, $s + t + u = p^2 + k^2 + p'^2 + k'^2$. This off-shell version of the constraint will become important when we discuss the so called *fixed-t dispersion relations*. When all our particles are on shell, this constraint takes the form $s + t + u = 2M^2 + 2\mu^2$.

The relation between the invariant t and the scattering angle θ depends on the Lorentz frame used. But Lorentz invariance requires scattering amplitudes to be functions of the invariants s, t, u . Often, the convenient frame is the so-called *centre of mass frame* where $\mathbf{p} + \mathbf{k} = 0$. In this frame, let us recall how the centre of mass energy W , the common magnitude K of $\mathbf{p}, \mathbf{k}, \mathbf{p}', \mathbf{k}'$, and, the scattering angle θ are related (all particles are on shell) to the invariants s, t :

$$\begin{aligned} W &= \sqrt{s} \quad K^2 = \frac{[s - (M + \mu)^2][s - (M - \mu)^2]}{4s} \\ t &\equiv -4\Delta^2 = -2K^2(1 - \cos \theta) \end{aligned} \quad (11.53)$$

Therefore, fixing the scattering angle will lead to a t (also called *momentum transfer*) that depends on s . A more invariant description is to consider scattering at fixed t , but different values of s . The fixed- t dispersion relations are the relevant structures to study in that context.

Depending on the objectives, Lorentz frames other than the centre of mass are also often used as we saw two different examples in Goldberger's as well as Symanzik's approaches. In what follows, we shall see a third of its kind in Salam's treatment of dispersion relations for non-forward scattering. Though Salam does not use a manifestly covariant approach, his dispersion relations are indeed an example of fixed- t dispersion relations.

Salam, in his paper *On Generalized Dispersion Relations* [9], attempts a proof of dispersion relations for non-forward scattering. He pretty much follows Goldberger's treatment of the forward scattering case but claims important generalizations for the non-forward case. It will turn out that by far his is the simplest approach to this problem. Since he mostly uses the same methodology as Goldberger, it will be pretty straightforward to explain Salam's paper. Unfortunately, it is marred by many typos, ambiguities and missing steps. In view of the extreme importance of this problem, we shall comment in detail on all these lacunae with the hope that some reader(s) will clean up things. To his credit, Salam attempts the most direct approach to this problem both physically and mathematically.

He too considers a field theory of neutral *Mesons* of mass μ interacting with *Nucleons* of mass M (he uses χ which we have changed to M). The meson field is $\phi(x)$. In Salam's case the mesons have no internal quantum numbers as in the case of Goldberger. As he considers no spin-flip for nucleons, the additional labels λ', λ used by Goldberger are also not there. Salam claims that the interaction Lagrangean is $j(x)\phi(x)$, and later uses $j(x)$ to signify the meson sources. But an interaction Lagrangean of his type does not yield $j(x)$ as the source when $j(x)$ itself depends on meson fields, so it is more consistent to treat $j(x)$ as the source without specifying any interaction Lagrangean.

Salam imposes the microcausality condition only on the sources

$$[j(x), j(y)] = 0 \quad (x - y)^2 < 0 \quad (11.54)$$

The metric convention used by Salam is $(- + + +)$ which is the same as the one used by Barton, Gasiorowicz, Bjorken and Drell, but opposite that of Goldberger.

Salam also starts with what Goldberger calls the *Feynman Matrix Element* (Salam too calls it the Feynman matrix element!) in his Eq. (2). Salam's paper suffers from a lot of notational clumsiness. He uses M to denote various kinds of amplitudes. To avoid confusion with the same symbol (but different meanings) used by Goldberger, Lehmann, etc., we shall denote Salam's Feynman matrix element by $\mathcal{M}_F$. Salam, like Goldberger, and unlike Symanzik, chooses to reduce the incoming and outgoing mesons. Writing this down explicitly:

$$\mathcal{M}_F(k, k'; p, p') = i \int d^4x d^4y e^{ik' \cdot x - ik \cdot y} \langle p' | T(j(x)j(0)) | p \rangle \quad (11.55)$$

where $T(j(x)j(y))$ is the T -product (Salam uses the notation $[j(x)j(y)]_+$ for it). Following Goldberger, he also introduces the *retarded matrix element*

$$\mathcal{M}(k, k'; p, p') = i \int dx dy e^{ik' \cdot x - ik \cdot y} \theta(x_0 - y_0) \langle p' | [j(x), j(y)] | p \rangle \quad (11.56)$$

Several remarks are in order at this stage:

- Neither of the Eqs. (11.55, 11.56) is what one gets from the LSZ reductions.
- Nor are they the Eqs. (2.4, 2.5) of Goldberger, which in fact are exactly what the LSZ reductions give, modulo some additional assumptions made by Goldberger, which we have already commented upon.
- Salam's expressions result after all equal time terms as well as terms with higher derivatives of delta functions are dropped. As per our discussions earlier, this is reasonable as long as dispersion relations are concerned.
- Goldberger had claimed (for $k_0 > 0$) the physical equivalence of his Eqs. (2.8, 2.9), which are the analogs of Salam's Eqs. (11.55, 11.56) only for forward scattering. Salam, on the other hand, claims equivalence (also for $k_0 > 0$) for non-forward scattering also. It should be appreciated that this equivalence is nothing more than the equivalence of the R -product and T -product representations we discussed in the context of the LSZ reductions. It holds rather generally. Therefore, Salam's assertion that this equivalence holds even for non-forward scattering is correct.
- From either of these forms it is clear that the form of microcausality used by Salam is sufficient.

In the next step, Salam writes down his Eq. (4), which is the version with a single space-time integration (essentially the T -matrix element). Though we have discussed how to go from two space-time integrations to just one, we will explicitly demonstrate this. It just amounts to using the space-time displacement operator P :

$$O(x) = e^{-iP \cdot x} O(0) e^{iP \cdot x} \quad (11.57)$$

Therefore

$$\begin{aligned}\mathcal{M}(k, k'; p, p') &= i \int dx dy e^{ik' \cdot x - ik \cdot y} \langle p' | e^{iP \cdot x} [j(0), j(y-x)]_{ret} e^{-iP \cdot x} | p \rangle \\ &= i \int dx dy e^{ik' \cdot x - ik \cdot y} e^{ip' \cdot x - ip \cdot x} \langle p' | [j(0), j(y-x)]_{ret} | p \rangle\end{aligned}\quad (11.58)$$

Introducing new variables of integration $x, x' = y - x$ in place of x, y , and performing the x -integration, it is easy to show that

$$\mathcal{M}(k, k'; p, p') = (2\pi)^4 \delta^{(4)}(p + k - p' - k') M_R(k; p, p') \quad (11.59)$$

where

$$M_R(k; p, p') = i \int d^4x \theta(-x_0) e^{-ik \cdot x} \langle p' | [j(0), j(x)] | p \rangle \quad (11.60)$$

which is exactly Eq. (4) of Salam. Modulo the equal time terms, this is also the same as Eq. (2.9) of Goldberger after indices $\alpha, \beta, \lambda, \lambda'$ were dropped, and due attention paid to the differing metric convention. This has prompted Salam to introduce the *advanced* matrix element

$$M_A(k; p, p') = i \int d^4x \theta(x_0) e^{-ik \cdot x} \langle p' | [j(0), j(x)] | p \rangle \quad (11.61)$$

Instead of translating by $-x$ in the above, one could have chosen to translate by $-y$, and one would have arrived at

$$\mathcal{M}(k, k'; p, p') = -(2\pi)^4 \delta^{(4)}(p + k - p' - k') M_A(-k'; p, p') \quad (11.62)$$

thus giving rise to the important identity

$$M_R(k; p, p') = -M_A(-k'; p, p') \quad (11.63)$$

Salam also notes the identities

$$M_{R,A}^*(k; p, p') = M_{R,A}(-k; p', p) \quad (11.64)$$

which are easily proved from the defining equations above.

Like Goldberger, Salam also finds it advantageous to work with a particular Lorentz frame to make further progress. Unlike Goldberger who chose the *Laboratory Frame* $\mathbf{p} = 0$ (since he had restricted his analysis to forward scattering only, this is all that matters), Salam chooses the rather unusual frame $\mathbf{p} + \mathbf{p}' = 0$ (which he incorrectly describes as *the frame in which nucleons are at rest*). He further chooses the direction for the nucleon momenta to be the 3-axis. Taking the magnitude of $\mathbf{p}$ to be P , the nucleon four-momenta are

$$p = \sqrt{P^2 + M^2}, 0, 0, P \quad p' = \sqrt{P^2 + M^2}, 0, 0, -P \quad (11.65)$$

By energy conservation $k_0 = k'_0$ and hence $|\mathbf{k}| = |\mathbf{k}'|$. Since the momentum components along 1,2-axes of both initial and final nucleons are zero, one immediately gets $k_1 = k'_1$; $k_2 = k'_2$. Conservation of momentum along the 3-axis gives $P + k_3 = -P + k'_3$ which implies $k_3 = -P$, and, $k'_3 = P$.

Thus the scattering amplitudes depend on P, k_0, k_1, k_2 . The matrix elements $\langle p' | [j(0), j(x)] | p \rangle$ can only depend on the invariants $(p + p') \cdot x, (p - p') \cdot x$, i.e. only on (P, x) . The entire k_0, k_1, k_2 dependence comes from the $e^{-ik \cdot x}$ factor. The Mandelstam invariants take on the values $t = (p - p')^2 = -4P^2, s = (p + k)^2 = (k_0 + \sqrt{P^2 + M^2})^2 - k_1^2 - k_2^2$. It is a nice feature of this frame that the *longitudinal* momentum P has the exact significance of quantifying the invariant momentum transfer t . On noting that $|\mathbf{k}|^2 = k_1^2 + k_2^2 + P^2$, it is seen that the Mandelstam variable s is $s(P, k_0) (k_0 = \sqrt{|\mathbf{k}|^2 + \mu^2})$. Another fact to note is that the *transverse momentum* of mesons both initially and finally satisfies $|\mathbf{k}_\perp|^2 = k_1^2 + k_2^2 = |\mathbf{k}|^2 - P^2$. Therefore it can always be traded in terms of $|\mathbf{k}|^2$ and P .

Salam's point is that even the problem of dispersion relations for non-forward scattering is mathematically of the same nature as the one treated by Goldberger for forward scattering, if P is held fixed. And, as just noted, this is the same as fixed- t scattering. He then on basically mimics all of Goldberger's steps. But of course, some crucial details are different!

As a first step in this direction, Salam introduces the analogs of $K_{\alpha\beta}^{(\pm)}(x)$ of the Goldberger analysis. Of course, with the internal labels for mesons α, β no longer present, only a single analog of the *symmetric* $K_{\alpha\beta}^{(-)}(x)$ remains. But now the scattering being non-forward, $p \neq p'$, Salam introduces two such functions:

$$\begin{aligned} J_1(P, x) &= \langle p | [j(0), j(x)] | p' \rangle - \langle p' | [j(0), j(x)] | p \rangle \\ J_2(P, x) &= i \{ \langle p | [j(0), j(x)] | p' \rangle + \langle p' | [j(0), j(x)] | p \rangle \} \end{aligned} \quad (11.66)$$

For forward scattering $J_1 = 0$, so indeed only one function remains as the analog of Goldberger's K-functions. Under the reasonable assumption that $j(x)$ are Hermitean, it is easy to show that both $J_1(P, x), J_2(P, x)$ are real. It will be useful to invert the above:

$$\langle p' | [j(0), j(x)] | p \rangle = -\frac{1}{2}(J_1(P, x) + i J_2(P, x)) \quad (11.67)$$

At this point, Salam makes the claim that $\theta(x) \langle p' | [j(0), j(x)] | p \rangle$ is an *even* function under $x_1 \rightarrow -x_1, x_2 \rightarrow -x_2$. In contrast, the matrix element of $K_{\alpha\beta}^{(-)}(x)$ was an *odd* function under $x \rightarrow -x$. But the forward nature of scattering was crucial in arriving at this. In fact, it is easy to show that

$$\langle p' \lambda' | K_{\alpha\beta}^{(-)}(-x) | p \lambda \rangle = -e^{i(p'-p)x} \langle p' \lambda' | K_{\alpha\beta}^{(-)}(x) | p \lambda \rangle \quad (11.68)$$

On the basis of this alleged evenness, Salam claims what is clearly a crucial result for his approach. That is

$$M_R(k_0, \mathbf{k}; p, p') = M_R(k_0, -\mathbf{k}; p', p) \quad (11.69)$$

It is tempting to compare this with Eq. (2.11a) of Goldberger (see the first of Eq. (11.18)); the two look very similar but with many important differences. Goldberger had invoked parity arguments in its derivation, but Salam makes no mention of such arguments. An important aspect of Salam's equation is that the order of p, p' is reversed from l.h.s to r.h.s. For forward scattering it may seem that this is undetectable. A natural generalization of this when additional quantum numbers like spin (labelled by λ in Goldberger's analysis) are included would be

$$M_R(k_0, \mathbf{k}; p, \lambda, p', \lambda') = M_R(k_0, -\mathbf{k}; p', \lambda', p, \lambda) \quad (11.70)$$

When restricted to the forward but incoherent scattering of Goldberger, i.e. $p' = p, \lambda' \neq \lambda$, this does not reproduce Goldberger's equation. Salam says he considers only the case of no spin-flip for nucleons, but it is not clear how that restriction would have entered his derivation of Eq. (11.69). It is tempting to consider the possibility that there is a typo in this equation and it should indeed be

$$M_R(k_0, \mathbf{k}; p, p') = M_R(k_0, -\mathbf{k}; p, p') \quad (11.71)$$

Indeed, Eq. (23) of Salam:

$$M_R = \frac{1}{2} [M_R(k_0, \mathbf{k}; p, p') + M_R(k_0, -\mathbf{k}; p, p')] \quad (11.72)$$

would strengthen that suspicion. But then, on using Salam's Eq. (4) (reproduced here in Eq. (11.60)) defining M_R , one would have got

$$M_R = \frac{1}{2} \int dx e^{-ik_0 x_0} \theta(-x_0) (J_2(P, x) - i J_1(P, x)) \cos \mathbf{k} \cdot \mathbf{x} \quad (11.73)$$

which is in complete disagreement with Salam's Eq. (23). Therefore we conclude that Eq. (11.69) is probably correct, and there is a typo in Salam's Eq. (23)! Then his Eq. (23) should be

$$\begin{aligned} M_R &= \frac{1}{2} [M_R(k_0, \mathbf{k}; p, p') + M_R(k_0, -\mathbf{k}; p', p)] \\ &= \frac{i}{2} \int dx e^{-ik_0 x_0} \theta(-x_0) [\langle p' | [j(0), j(x)] | p \rangle e^{-i\mathbf{k} \cdot \mathbf{x}} + \langle p | [j(0), j(x)] | p' \rangle e^{i\mathbf{k} \cdot \mathbf{x}}] \\ &= \frac{1}{2} \int dx e^{-ik_0 x_0} \theta(-x_0) [J_1(P, x) \sin \mathbf{k} \cdot \mathbf{x} + J_2(P, x) \cos \mathbf{k} \cdot \mathbf{x}] \end{aligned} \quad (11.74)$$

This indeed looks almost the same as Salam's expression. But he has a $\theta(x_0)$ in place of the $\theta(-x_0)$ above. It is not clear how he gets this. Again, it does not seem to be a typo as equations he deduces from this continue to have $\theta(x_0)$!

Let us get back to the mysterious Eq. (11.69), as this is the most crucial ingredient to his derivations. Though he does not show this, Salam connects this to his claimed *evenness* of $\theta(x_0) \langle p' | [j(0), j(x)] | p \rangle$ under $x_1 \rightarrow -x_1, x_2 \rightarrow -x_2$

while x_0, x_3 are kept unchanged (we shall henceforth call this *transverse parity* transformation.). In Goldberger's treatment, which looked at oddness under $x \rightarrow -x$, the demonstration consisted of *four-displacement* by $-x$ so $[j(0), j(x)] \rightarrow e^{-iPx} [j(-x), j(0)] e^{-iPx}$. But what we have called transverse parity cannot be effected through displacements alone, as can be easily checked. It can however be realized as $x \rightarrow -x$, which can be effected through the displacements, followed by $x_3 \rightarrow -x_3$ which is pure rotation, and, $x_0 \rightarrow -x_0$. This last operation is *Time Reversal* and it has to be represented by *Anti-unitary* transformations, which necessarily involves interchanging initial and final states. This may explain why p, p' get interchanged in Eq. (11.69).

Following Goldberger, Salam also introduces the dispersive and absorptive parts $D(k_0, P)$ and $A(k_0, P)$, defining them to be (see his Eqs. (6, 7)):

$$D(k; p, p') = \frac{(M_R - M_A)}{2} \quad A(k; p, p') = \frac{(M_R + M_A)}{2i} \quad (11.75)$$

He then deduces these from his Eq. (23) (our Eq. (11.74)) to be:

$$\begin{aligned} D(k_0, P) &= \frac{1}{2} \int d^4x \theta(x) \cos k_0 x_0 (\cos k_1 x_1 + k_2 x_2) \\ &\quad \cdot \{J_1(P, x) \sin Px_3 + J_2(P, x) \cos Px_3\} \\ A(k_0, P) &= -\frac{1}{2} \int d^4x \theta(x) \sin k_0 x_0 (\cos k_1 x_1 + k_2 x_2) \\ &\quad \cdot \{J_1(P, x) \sin Px_3 + J_2(P, x) \cos Px_3\} \end{aligned} \quad (11.76)$$

There are several comments in order at this stage: i) Salam continues to claim factors of $\theta(x)$ though his expression for M_R involved a $\theta(-x_0)$, ii) more seriously, irrespective of whether M_R comes with a $\theta(-x_0)$ or $\theta(x_0)$ (and M_A other way around), $D(k_0, P)$ being proportional to $M_R - M_A$ should involve $\epsilon(x_0)$ in its integrand; likewise, $A(k_0, P)$ being proportional to $M_R + M_A$ should have unity in place of the theta function. This was so with Goldberger's D and A . In fact, that should be a fairly general feature, so Salam's expressions seem *incorrect*. Funnily enough, Salam's own Eq. (29) for A , when he is trying to relate A to observables, does not have any $\theta(x_0)$!, iii) in going from Eq. (11.74) to these expressions, one has to use (on noting $\mathbf{k} \cdot \mathbf{x} = -Px_3 + k_1 x_1 + k_2 x_2$) the following trigonometric relations:

$$\begin{aligned} \sin \mathbf{k} \cdot \mathbf{x} &= \sin(k_1 x_1 + k_2 x_2) \cos Px_3 - \cos(k_1 x_1 + k_2 x_2) \sin Px_3 \\ \cos \mathbf{k} \cdot \mathbf{x} &= \cos(k_1 x_1 + k_2 x_2) \cos Px_3 + \sin(k_1 x_1 + k_2 x_2) \sin Px_3 \end{aligned} \quad (11.77)$$

But the $\sin(k_1 x_1 + k_2 x_2)$ pieces are missing in Salam's expressions for D and A . These terms are *odd* under transverse parity, so their contributions to the integrals can vanish if indeed $\langle p' | [j(0), j(x)] | p \rangle$ is *even* under this transformation, as claimed by Salam (unfortunately, without proof!). Even then, one only gets the combination $-J_1 \sin Px_3 + J_2 \cos Px_3$ and not the combination $J_1 \sin Px_3 + J_2 \cos Px_3$ as claimed by Salam. How seriously this will affect the final conclusions is not clear.

Because of the rotational invariance (in the x_1, x_2 -plane) of the $dx_1 dx_2$ measure, the result of the x_1, x_2 integrations will only produce functions of $k_1^2 + k_2^2$ which can be rearranged as functions of (P, k_0) . Thus the functions D and A are indeed functions of (k_0, P) only. The last (almost) part of Salam's proof is to show, by direct evaluation, that

$$\frac{D(\alpha, P) - D(\beta, P)}{\alpha^2 - \beta^2} = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{k_0 A(k_0, P) dk_0}{(k_0^2 - \alpha^2)(k_0^2 - \beta^2)} \quad (11.78)$$

Salam says he is following Goldberger here (among other things, the interchange of the order of k_0 and x -integrations) and his treatment is subject to the same limitations as what Goldberger's would be. He also says that it is at this stage that the micro-causality condition of Eq. (11.54) is invoked but does not give any details. If one recalls our earlier discussion of how Goldberger handled similar integrations, we had noted the crucial replacement of $\sin \omega x_0, \cos \omega x_0$ respectively by $-ie^{i\omega x_0}, e^{i\omega x_0}$. Goldberger's explanations based on causality for this replacement were found not very clear, and we gave a more straightforward explanation which heavily depended on the oddness/evenness of the $K_{\alpha\beta}$'s as well as the occurrence only of $\cos \mathbf{k} \cdot \mathbf{x}$ in the integrand. In Salam's case neither of these holds. Therefore it is very important to deconstruct this crucial part of Salam's paper. Another important factor in Goldberger's treatment is the restriction of the x_0 integration to only positive values. But Salam mysteriously switches from his earlier $\theta(-x_0)$ to $\theta(x_0)$, so this part is also not at all clear.

Lastly, Salam also attempts relating $A(k_0, P)$ to observables to give dispersion relations an observable content, as did Goldberger and Symanzik. We shall not go into the details of this. The expression for $A(k_0, P)$ that Salam uses here in his Eq. (29) is indeed without any θ -function, as it should be, but contradicting his earlier Eq. (25) (reproduced here in Eq. (11.76)).

Though we have been very critical of various shortcomings in Salam's paper, it still offers a way to proving fixed-t dispersion relations that is transparent both mathematically and physically, on par with Goldberger's treatment of the forward scattering case. But it obviously needs a lot of cleaning up. Salam talks of a second paper taking into account spin-flip for nucleons. Hopefully, that paper may shed light on the various inadequacies highlighted here.

11.6 Fixed-t Dispersion Relations: Lehmann and Sommer

Now we turn to perhaps the most rigorous approaches to the problem of fixed-t dispersion relations. The pioneering works are those of Bogoliubov, Medvedev, Polivanov [12], of Bremermann, Oehme and Taylor [13], and, of Bros, Epstein and Glaser [14]. These approaches are very technical involving many aspects of the *Functions of Several Complex Variables*. The reasons for the proliferation of complex variables are essentially twofold: a) the invariants (s, t) have to be complexified, and, b) all the *mass invariants* p^2, p'^2, k^2, k'^2 also have to be analytically extended. The

need for b) is far from obvious, but marks the high point of the technical analyses mentioned above.

We shall not explain these details as they are very terse and algebraically tedious. Also, the theory of functions of several complex variables is much harder and highly counter-intuitive. Instead, we shall follow the description given by Lehmann in his lectures [15] which is rather transparent, but there are some important gaps. Where that happens, we shall supplement it with suitable references to the excellent review by Gustav Sommer [16]. The main technical problem is the determination of various *Analyticity Domains*. We have already given a flavour of this problem in some detail. The Sommer review gives a lot of important technical details as to how such domains are actually computed. It also has a nice discussion of important theorems concerning functions of several complex variables that are relevant.

As a warm-up let us revisit our earlier discussion of the so-called Lehmann Ellipses, both small (SLE) as well as the large (LLE) for the imaginary part of the scattering amplitude. There, it was shown how the regularity(analyticity) of the scattering amplitude as a function of $z = \cos \theta$, θ being the scattering angle, could be extended from part of the real line $-1 \leq z \leq +1$ to the Lehmann Ellipses in complex z -plane. That followed directly from the LSZ formalism via the integral representation for scattering amplitudes. The latter was a consequence of the Jost-Lehmann representation [17] as extended by Dyson [18] for *unequal mass* scatterings. Except for some very mild assumptions like *single particle stability*, *vacuum stability* and absence of stable bound states, the treatments are very general, without making additional simplifying assumptions (however reasonable they may appear to be) as, for example, made by Goldberger, Symanzik and Salam.

As we have covered all this in great detail, it suffices to give an overview of the essentials. The starting point is the scattering amplitude for the scattering of two particles into two particles. We refer to Lehmann's 1958 article [19] for details, which we have already discussed threadbare earlier. The system considered by him are spin-less mesons of mass μ characterized by field $\phi(x)$, and, spin-less charged nucleons of mass M , characterized by fields $\psi(x)$. The T -matrix representation used by Lehmann, when both the final particles are reduced is

$$T = -i \int d^4x e^{i \frac{(k' - p')}{2} \cdot x} \langle 0 | R'(\phi(\frac{x}{2}) \psi(-\frac{x}{2})) | pk, in \rangle \quad (11.79)$$

After imposing the requirements of causality, and applying the Jost-Lehmann-Dyson integral representation, Lehmann obtains

$$T = \frac{1}{2\pi} \int \dots \int d^4u d\kappa^2 \frac{\bar{\phi}(u, \kappa^2 p, k)}{(\frac{(k' - p')}{2} - u)^2 - \kappa^2} \quad (11.80)$$

In the centre of mass frame, defined by $\mathbf{p} + \mathbf{k} = 0$, this integral representation can be recast(see earlier chapters) as

$$T(W, \cos \theta) = \int_{x_{\min}(W)}^{\infty} dx \int_0^{2\pi} d\alpha \frac{\bar{\phi}(x, \cos \alpha, W)}{x - \cos(\theta - \alpha)} \quad (11.81)$$

where W , $\cos \theta$ are the total energy and scattering angle in the centre of mass frame. As already explained,

$$x_{\min}(W) = \left\{ 1 + \frac{(m_1^2 - \mu^2)((m_2^2 - M^2))}{K^2 (W^2 - (m_1 - m_2)^2)} \right\}^{\frac{1}{2}} \quad (11.82)$$

with K being the magnitude of the centre of mass momentum:

$$K^2 = \frac{[W^2 - (M + \mu)^2][W^2 - (M - \mu)^2]}{4W^2} \quad (11.83)$$

It should be carefully noted that in the above, all particles were *on mass shell*. It should also be remembered that m_1 is the lowest mass of intermediate states with only mesons, and, m_2 that of the intermediate states with one nucleon and a meson. In the present case $m_1 = 3\mu$ when certain selection rules are imposed is 2μ otherwise. Likewise $m_2 = M + \mu$. It is also usually assumed that $m_2 > m_1$, which presently would imply $M > 2\mu$. Many aspects of analyticity of scattering amplitudes, and so also of dispersion relations, depend on masses in a critical way; this is one of the reasons for the difficulty in arriving at general proofs.

As already discussed earlier, the great advantage of the representation of Eq. (11.81) is that the entire dependence on the scattering angle θ , or rather its cosine $z = \cos \theta$, is explicitly contained in the denominator. This immediately answers how the domain of analyticity in z can be extended from the physical domain $-1 \leq z \leq +1$. The fact that $x_{\min}(W) > 1$ is critical. Since singularities of the integral can only occur when this denominator vanishes, the analyticity domain for z gets extended to

$$z = \cos \theta = x \cos \alpha \pm i \sqrt{x^2 - 1} \sin \alpha \quad (11.84)$$

In other words, the scattering amplitude is an *analytic* function of complex- z which is regular inside an ellipse in the complex z -plane centred at the origin, with axes $x_{\min}, \sqrt{x_{\min}^2 - 1}$, and, foci at $z = \pm 1$. Likewise, $\text{Im } T(W, \cos \theta)$ is regular in a *larger* ellipse with axes $y_{\min} = 2x_{\min}^2 - 1; \sqrt{y_{\min}^2 - 1} = 2x_{\min}\sqrt{x_{\min}^2 - 1}$, also centred at the origin, and foci at $z = \pm 1$.

Though we discussed all this in great detail in previous chapters, we are drawing attention to them to set the stage for the discussion of larger analyticity issues that fixed- t dispersion relations demand. It is often more useful to recast analyticity in z in terms of the invariant $t = -2K^2(1 - \cos \theta)$.

Therefore, what the Lehmann ellipses demonstrate is how to expand the analyticity domain in complex- t , when the invariant s is fixed, and importantly, when all particles are on mass shell, i.e. $p^2 = p'^2 = M^2; k^2 = k'^2 = \mu^2$. The extended analyticity domains required for the validity of fixed- t dispersion relations are totally different; they are domains of maximum analyticity in complex s -plane, for fixed t (fixed for physical values). Superficially, these two problems appear to be different and unconnected. But we will soon see a very intimate connection between them.

As stated by Sommer (immediately after the beginning of his Sect. 4.2 on Large Lehmann Ellipses), provided t lies in the Large Lehmann Ellipse, analyticities in mass variables $\xi_1 = p^2$, $\xi_2 = k^2$, $\xi_3 = p'^2$, $\xi_4 = k'^2$, that too in a *small* neighbourhood of their mass shell values $\xi_1 = \xi_3 = M^2$, $\xi_2 = \xi_4 = \mu^2$, which are also consequences of the Large Lehmann Ellipse analyticity, allow a relatively simple proof of the fixed- t dispersion relations.

Therefore the big surprise is the need for analytically extending amplitudes in the mass variables. If we momentarily change notation such that $p_1 = p$, $p_2 = k$, $p_3 = p'$, $p_4 = k'$, then the *invariant matrix* $p_i \cdot p_j$ has as its diagonal elements the mass variables p_i^2 , and the other invariants s, t, u are built out of these mass variables as well as the off-diagonal elements. The overall energy-momentum conservation $p + k = p' + k'$ introduces dependencies among the invariants:

$$s + t + u = \xi_1 + \xi_2 + \xi_3 + \xi_4 \quad (11.85)$$

Thus, the problem of determining the analyticity domains of scattering amplitudes, in its fullest generality, is a problem in the theory of *six* complex variables. However, many essential results can still be obtained by restricting these to fewer in number. We will follow Lehmann's lectures to understand the broad features. The interested reader should follow the Sommer review for greater technical details and completeness.

Lehmann chooses a different LSZ reduction than in Eq. (11.79) as a starting point for the discussion:

$$T = -i \int d^4x e^{i \frac{(k+k')}{2} \cdot x} \langle p' | R'(\phi(\frac{x}{2}) \phi(-\frac{x}{2})) | p \rangle \quad (11.86)$$

The main point raised by Lehmann is that the above representation need not be restricted only to cases when all momenta are on shell. It can be taken to *define* T for at least all *real* values of k, k' . In fact, Sommer treats the even more general case of when momenta are complex too, as this will be found essential. Nevertheless, the momenta are always restricted by $p + k = p' + k'$. Here, the mesons are the ones that have been reduced. Not only that, Lehmann chooses to keep both the nucleons on shell, i.e. $\xi_1 = \xi_3 = M^2$, while the off-shell values of ξ_2, ξ_4 are chosen to be the same, i.e. $\xi_2 = \xi_4 = \xi$. Now, the amplitude T is a function of three invariants:

$$s = (p + k)^2; \quad \Delta^2 = -\frac{(p - p')^2}{4} = -\frac{t}{4}; \quad \xi = k^2 = k'^2 \quad (11.87)$$

Lehmann additionally introduces the invariant

$$\omega = \frac{(k + k') \cdot (p + p')}{2\sqrt{(p + p')^2}} \quad (11.88)$$

As a result of the mesons being off-shell the earlier expressions for centre of mass variables W and K have to be modified:

$$W^2 = 2\omega\sqrt{\Delta^2 + M^2} + 2\Delta^2 + M^2 + \xi; \quad K^2 = \frac{(W^2 + M^2 - \xi)^2 - 4M^2W^2}{4W^2} \quad (11.89)$$

Thus, the variable ω can be taken in lieu of W , and T as $T(\omega, \Delta^2, \xi)$.

The physical scattering amplitude, which is a function of (s, t) , equivalently (ω, Δ^2) is to be understood as the boundary value

$$T(\omega, \Delta^2) = T(\omega, \Delta^2, \xi = \mu^2) \quad (11.90)$$

The scattering angle is given by

$$\cos \theta = 1 - \frac{2\Delta^2}{K^2(\xi = \mu^2)} \quad (11.91)$$

Following Lehmann, we introduce the special Lorentz frame $\mathbf{p} + \mathbf{p}' = 0$; this frame has already been introduced before. It immediately follows that $\Delta^2 = -\frac{(\mathbf{p}-\mathbf{p}')^2}{4} = |\mathbf{p}|^2$. It is straightforward to verify the following results given by Lehmann:

$$\begin{aligned} p &= (\sqrt{\Delta^2 + M^2}, \mathbf{p}); & k &= (\omega, -\mathbf{p} + \mathbf{e}\sqrt{\omega^2 - \Delta^2 - \xi}) \\ p' &= (\sqrt{\Delta^2 + M^2}, -\mathbf{p}); & k' &= (\omega, \mathbf{p} + \mathbf{e}\sqrt{\omega^2 - \Delta^2 - \xi}) \end{aligned} \quad (11.92)$$

here $\mathbf{e}$ is a unit-vector satisfying $\mathbf{e} \cdot \mathbf{p} = 0$. Now the significance of the ω -invariant stands out; in the special frame chosen, it is the meson energy both initially and finally. In the physical region, $\omega^2 > \Delta^2 + \mu^2$. Writing $T(\omega, \Delta^2, \xi)$ in this frame:

$$T(\omega, \Delta^2, \xi) = - \int d^4x e^{i\omega x_0 - i\mathbf{e} \cdot \mathbf{x} \sqrt{\omega^2 - \Delta^2 - \xi}} \langle p' | R'(\phi(\frac{x}{2})\phi(-\frac{x}{2})) | p \rangle \quad (11.93)$$

After imposing microcausality, i.e. $x_0 \geq |\mathbf{x}|$, the question of determining the domains of analyticity in complex- ω (and, importantly, in ξ) boils down to determining the region

$$\text{Im } \omega > |\text{Im } \sqrt{\omega^2 - \Delta^2 - \xi}| \quad (11.94)$$

A similar problem arose in Capps and Takeda's analysis earlier, and in related forms in the analyses of Goldberger, Symanzik and Salam. We had also remarked that none of these had satisfactorily addressed this central issue. The added complication now is the variable ξ , and that also has to be complexified so one has to address the much harder problem of analyticity in both ω, ξ . While Symanzik writes down the extended analyticity domain in his Eq. (47a), the steps leading to that are not at all obvious.

On the other hand, Sommer gives a very systematic treatment of this issue. It is algebraically tedious, and not particularly illuminating, so we shall be content with

mentioning the final results. However, we will explain the major strategy adopted by Sommer. He extends the earlier, on shell, derivation of the Lehmann Ellipses to take care of off-shell cases. In the earlier analysis of on-shell scattering, the axis of the Large Lehmann Ellipse y_0 depended on s , the masses as well as the thresholds. Let us start with Sommer expression, still for the on-shell case, but when all masses are different. The axes of the small Lehmann Ellipses are given by

$$x_{\min,34}(s) = \left\{ 1 + \frac{(M_3^2 - m_3^2)(M_4^2 - m_4^2)}{K_2^2(s - (M_3 - M_4)^2)} \right\}^{\frac{1}{2}} \quad (11.95)$$

Here m_1, m_2, m_3, m_4 are the masses of the particles, while M_i are the lowest masses of intermediate states with quantum numbers pertaining to particles of mass m_i . We have introduced the notation $x_{\min,34}$ to denote the case where the particles in the final state are 3, 4. K_2 is the common magnitude of the momenta in the final state, $K_2 = |\mathbf{p}'| = |\mathbf{k}'|$. Likewise, K_1 is the common magnitude for the initial state. We write these down explicitly:

$$\begin{aligned} K_1^2 &= \frac{(s - (m_1 + m_2)^2)(s - (m_1 - m_2)^2)}{4s} \\ K_2^2 &= \frac{(s - (m_3 + m_4)^2)(s - (m_3 - m_4)^2)}{4s} \end{aligned} \quad (11.96)$$

It is important to note that the small Lehmann ellipse is only dependent on the final state masses. The expressions for the axes of the Large Lehmann Ellipses are given by

$$y_0(s) = x_{\min,34} x_{\min,12} + \sqrt{x_{\min,34}^2 - 1} \sqrt{x_{\min,1,2}^2 - 1} \quad (11.97)$$

Unlike the small Lehman Ellipse, the large Lehmann Ellipse depends on both the particles in the initial state, as well as those in the final state. Sommer goes beyond the on-shell cases. From the derivation of the Lehmann Ellipses, it is clear that the case where all particles are off-shell is obtained by replacing m_i^2 by ξ_i in all expressions (including K_1, K_2). The threshold masses M_i are unaffected. We write down the expression for the Large Lehmann Ellipse when all four particles are off-shell:

$$\begin{aligned} y_0(s) &= \sqrt{1 + \frac{(M_3^2 - \xi_3)(M_4^2 - \xi_4)}{K_2^2(s - (M_3 - M_4)^2)}} \sqrt{1 + \frac{(M_1^2 - \xi_1)(M_2^2 - \xi_2)}{K_2^2(s - (M_1 - M_2)^2)}} \\ &\quad + \sqrt{\frac{(M_3^2 - \xi_3)(M_4^2 - \xi_4)}{K_2^2(s - (M_3 - M_4)^2)}} \sqrt{1 + \frac{(M_1^2 - \xi_1)(M_2^2 - \xi_2)}{K_2^2(s - (M_1 - M_2)^2)}} \end{aligned} \quad (11.98)$$

With (the expressions have been recast to bring out the symmetry in masses explicitly)

$$K_1^2 = \frac{(s - \xi_1 - \xi_2)^2 - 4\xi_1\xi_2}{4s} \quad K_2^2 = \frac{(s - \xi_3 - \xi_4)^2 - 4\xi_3\xi_4}{4s} \quad (11.99)$$

To establish fixed- t dispersion relations, it is sufficient to take only the mesons off-shell but leave the nucleons (of mass M) on shell both initially and finally, i.e. $\xi_1 = \xi_3 = M^2$. Furthermore, it also suffices to take the off-shell values for the initial and final mesons to be the same, i.e. $\xi_2 = \xi_4 = \xi$. Then this becomes a case of *elastic scattering* for which $y_0(s) = 2x_{\min}^2(s) - 1$. Also $K_1 = K_2$ in that case. The relevant expression for $y_0(s)$ becomes

$$y_0(s) = 1 + \frac{2(M_1^2 - M^2)(M_2^2 - \xi)}{K^2(\xi)(s - (M_1 - M_2)^2)} \quad K^2(\xi) = \frac{(s + M^2 - \xi)^2 - 4sM^2}{4s} \quad (11.100)$$

This expression is the starting point for the determination of the relevant analyticity domains, as discussed in detail by Sommer in his Sect. 5.1. We skip the details and present only the essential results. He arrives at a variety of analytical domains. For fixed- t dispersion relations, one is interested in analyticity domains that is the entire upper half of complex s -plane. One such, discussed in detail by Sommer is his domain X_+ described in his Eq. (5.9). This domain is characterized by $\text{Im } s > \text{Im } \xi$, and for *real* ξ , it contains the entire upper half of complex s -plane. From Sommer's Eq. (5.9), it is also clear that $\xi < \frac{t}{4}$, and that t should lie in the Large Lehmann Ellipse. Another such domain is X_0 described in Eq. (5.12) of Sommer. Both these domains coincide with the domain given by Lehmann when the condition $\xi < -\Delta^2$ (for real ξ).

As per the summary given by Sommer on the bottom of p. 620, the corresponding advanced function is analytic in the *lower* half complex s -plane, and together they yield (Eq. (5.13) of Sommer):

$$T(s, t, \xi) = \frac{1}{\pi} \int_{M_1^2}^{\infty} ds' \frac{A_s(s', t, \xi)}{s' - s} + \frac{1}{\pi} \int_{M_2^2}^{\infty} du' \frac{A_u(u', t, \xi)}{u' - u} \quad (11.101)$$

Here A_s, A_u are the s -channel and u -channel discontinuities. The convergence of these integrals is assumed. If they do not, suitable subtractions will be necessary.

Lehmann too writes down similar relations, though only for the retarded part:

$$T(\omega, \Delta^2, \xi) = \frac{1}{\pi} \int_{-\infty}^{\infty} d\omega' \frac{\text{Im } T(\omega', \Delta^2, \xi)}{\omega' - \omega} \quad (11.102)$$

Here too convergence of integrals is tacitly assumed.

Though this looks like fixed- t dispersion relations, it is so for the off-shell $T(\omega, \Delta^2, \xi)$. To obtain the analogous dispersion relations for the *physical* amplitudes $T(\omega, \Delta^2, \xi = \mu^2)$, one has to analytically continue the above to the physical point $\xi = \mu^2$. This is again rather technical. One has to follow Bogoliubov's work for this [12]. Sommer has given the detailed arguments. For us, it suffices to say that such a continuation can indeed be carried out. We mention the conditions under which such a continuation exists. For this, consider the off-shell generalization

$$\text{Im } T(\omega, \Delta^2, \xi) = \frac{1}{2} \{M(\omega, \Delta^2, \xi) - M(-\omega, \Delta^2, \xi)\} \quad (11.103)$$

Bogoliubov et al. condition for being able to analytically continue to the physical point $\xi = \mu^2$ is: $M(\omega, \Delta^2, \xi)$ is an analytic function of ξ which is regular for $\text{Re } \xi \leq \mu^2$ in a *small neighbourhood* of the real axis, i.e. $|\text{Im } \xi| < \delta$.

Then one finally has the fixed-t dispersion relation:

$$T(\omega, \Delta^2) = \frac{1}{\pi} \int_{-\infty}^{\infty} d\omega' \frac{\text{Im } T(\omega', \Delta^2)}{\omega' - \omega} \quad (11.104)$$

11.7 Mandelstam Double Spectral Representation

No discussion of analyticity properties of scattering amplitudes, and, of dispersion relations can be complete without at least a brief mention of the *Mandelstam Double Spectral Representation*, often called *Mandelstam Representation* [20,21]. With his deep mastery of the subject and his mathematical powers, Mandelstam proposed that the scattering amplitude $A(s, t)$ for the elastic scattering of two spin-less mesons (or the invariant functions occurring in more complicated situations with spin and internal quantum numbers) should be an analytical function of two complex variables s, t over their entire complex planes except for the poles and branch cuts as demanded by unitarity. It should be remembered that these poles and branch cuts will be the totality of such singularities for the s-channel, t-channel and u-channel *physical processes*. The inspiration for the double spectral form came to Mandelstam from his many investigations into analyticity based on perturbative quantum field theory calculations combined with his deep conviction that amplitudes for all physical processes must be seen as appropriate boundary values of such analytic functions. As stated by Goddard [22] the Mandelstam Representation provided the inspiration for much of the study of analytic properties of scattering amplitudes.

$$\begin{aligned} A(s, t) &= \frac{1}{\pi^2} \int ds' \int dt' \frac{\rho_{st}(s', t')}{(s' - s)(t' - t)} + \frac{1}{\pi^2} \int dt' \int du' \frac{\rho_{tu}(t', u')}{(t' - t)(u' - u)} \\ &\leq + \frac{1}{\pi^2} \int du' \int ds' \frac{\rho_{us}(u', s')}{(u' - s)(s' - t)} \end{aligned} \quad (11.105)$$

All the integrals are taken along the respective real axes. The limits of integration, not explicitly shown above are: (s_0, ∞) for s' -integration, etc. s_0 is determined by the beginning of the branch cut due to s-channel unitarity, etc. The region of integration is determined by the regions where the spectral functions $\rho_{st}, \rho_{tu}, \rho_{us}$ are non-vanishing (see the book *Complex Angular Momentum and Particle Physics* by Squires [23], and, the pedagogically excellent treatment of the Mandelstam Representation by Radha [24] for more details). As mentioned earlier, these spectral functions must be such that the boundary value of $A(s, t)$ for the physical regions, i.e. appropriate regions in s, t plane where the variables take their physically permissible values like $s \geq (M + \mu)^2, -1 \leq z \leq +1$ for s-channel processes, must be the physical scattering amplitudes. Additionally, they must be such as to be consistent

with crossing symmetry, i.e. relating $A(s, t)$ in certain unphysical regions to physical amplitudes for the crossed processes.

These requirements make a determination of the spectral functions highly non-trivial. In addition, the behaviour of the spectral functions for very large values of their arguments must be such as to make the Mandelstam Representations *convergent*. Else, “subtractions” have to be made sufficiently many times to guarantee convergence. This can in principle be done at the double spectral representation level itself (see, for example, Marcel Froissart’s seminal paper [25]) but we will mention alternate strategies shortly. We will also see in the next chapter that this issue of the asymptotic behaviour is elegantly resolved by the so-called *Regge Poles*, whose existence in itself can be linked with the validity of the Mandelstam Representation.

As emphasized by Gasiorowicz (see Chap. 22 of [8]), the spectral functions could be composed of delta functions and their derivatives. As Froissart puts it [25], they are in general *tempered distributions*. It is instructive to see how the Mandelstam representation is invoked in practice. We refer the reader to Chap. 23 of the book by Gasiorowicz [8] wherein he treats Pion-Nucleon scattering. Pions, which are spin-less, have *Isotopic Spin* $I = 1$; on the other hand Nucleons have both spin and Isospin of $\frac{1}{2}$. It can be shown that all amplitudes can be expressed in terms of $A^{(\pm)}(s, t, u)$, $B^{(\pm)}(s, t, u)$ (see [8] for details). It is to these *invariant amplitudes* that the double spectral representation is applied.

In the s -channel, there is a *nucleon pole* at $s = M^2$, and a branch cut starting at $s_0 = (M + \mu)^2$. Likewise, the u -channel also has a single nucleon pole, and a cut also starting at $u_0 = (M + \mu)^2$. In the t -channel, the cut starts at $t_0 = 4\mu^2$. In Gasiorowicz’s treatment, no single particle poles are introduced in the t -channel. This is a matter of phenomenological detail; for example, a ρ -pole could have been introduced in the $I = 1$ channel. It is customary to separate out the contributions of these poles and write the Mandelstam Representation only for the remainder, as indeed has been done in [8].

It is immediately obvious that the double spectral representation can be rewritten in the single spectral form, and in particular as fixed- t dispersion relations:

$$A(s, t) = \frac{1}{\pi} \int dt' \frac{A_t(s, t')}{t' - t} + \frac{1}{\pi} \int du' \frac{A_u(s, u')}{u' - u} \quad (11.106)$$

where

$$\begin{aligned} A_t(s, t') &= \frac{1}{\pi} \int ds' \frac{\rho_{st}(s', t')}{s' - s} + \frac{1}{\pi} \int du' \frac{\rho_{tu}(s', u')}{u' - u} \\ A_u(s, u') &= \frac{1}{\pi} \int ds' \frac{\rho_{us}(s', u')}{s' - s} + \frac{1}{\pi} \int dt' \frac{\rho_{tu}(t', u')}{t' - t} \end{aligned} \quad (11.107)$$

Thus, the fixed- t dispersion relations follow as a consequence of the Mandelstam Representation. If one examines closely the role played by the double spectral representation in some important applications like the Froissart Bound [25], or the analytical continuation to complex angular momentum (Regge Theory), one sees that it is only in providing the fixed- t dispersion relations.

But we have already seen that fixed- t dispersion relations have been proved independently of the Mandelstam Representation. Actually, it is the Mandelstam Representation that has defied all attempts to prove it. Though Mandelstam himself used many perturbative quantum field theoretic calculations to gain intuition about the singularity structure of scattering amplitudes, perturbative QFT in itself was unable to establish the Mandelstam Representation.

Quite clearly, amplitudes in perturbative QFT do show the double spectral form. For example, the light-by-light scattering amplitude is one such. A very nice treatment of this can be found in Volume II of Julian Schwinger's *Particles, Sources And Fields* (Sect. 4.10) [26]. The lecture notes of Radha [24] also provide detailed calculation of such *Box Diagrams* from the point of view of the Mandelstam Representation. Among other things, it brings out the sensitive dependence of the analytic structures on masses.

Both Landshoff and collaborators [27], as well as Eden [28] had carried out extensive, all-order calculations in perturbation theory, and had claimed proofs of the Mandelstam Representation. But both the groups, on further analyses, came to the realization that their proofs were *incomplete* as they found isolated real singularities that spoiled the proof. This was acknowledged in a joint work of all the authors [29]. For a detailed understanding of the various technicalities, the reader is encouraged to consult the book *The Analytic S-matrix* [30]. The Mandelstam Representation remains unproven to this day!

References

1. J. Tiomno, W. Schutze, Phys. Rev. **83**, 249 (1951)
2. N.G. van Kampen, Phys. Rev. **89**, 1072 (1953)
3. N.G. van Kampen, Phys. Rev. **91**, 1267 (1953)
4. J.S. Toll, *Dispersion Relations and Its Applications to Problems Involving Electron Pairs*. Princeton University Thesis (1951)
5. J.S. Toll, Phys. Rev. **104**, 1760 (1956)
6. M.L. Goldberger, Phys. Rev. **99**, 979 (1955)
7. M.L. Goldberger, Phys. Rev. **97**, 508 (1955)
8. S. Gasiorowicz, *Elementary Particle Physics*. Wiley
9. A. Salam, Nuovo Cimento **3**, 424 (1956)
10. R.H. Capps, G. Takeda, Phys. Rev. **103**, 1878 (1956)
11. K. Symanzik, Phys. Rev. **105**, 743 (1957)
12. N.N. Bogoliubov, B. Medvedev, M. Polivanov, Fortschritte der Phys. **6**, 169 (1958)
13. H.J. Bremermann, R. Oehme, J.G. Taylor, Phys. Rev. **109**, 2178 (1958)
14. J. Bros, H. Epstein, V. Glaser, Comm. Math. Phys. **1**, 240 (1965)
15. H. Lehmann, Supp. Nuovo Cimento **14**, 153 (1959)
16. Gustav Sommer, Fortschritte der Phys. **18**, 577 (1970)
17. R. Jost, H. Lehmann, Nuovo Cimento **5**, 1598 (1957)
18. F.J. Dyson, Phys. Rev. **110**, 1460 (1958)
19. H. Lehmann, Nuovo Cimento 10 No 4, p. 579 (1958)
20. S. Mandelstam, Rep. Prog. Phys. **25**, 99 (1962)
21. S. Mandelstam, Phys. Rev. **112**, 1334 (1958)
22. P. Goddard, in *Memorial Volume for Stanley Mandelstam*, ed by N. Burkovits (World Scientific, 2017)

23. E.J. Squires, *Complex Angular Momentum and Particle Physics*. W.A. Benjamin Publishers
24. T.K. Radha, K. Venkatesan, *Lectures on Mandelstam Representation*, Matscience Report, vol. 6 (1963)
25. M. Froissart, Phys. Rev. **123**, 1053 (1961)
26. J. Schwinger, *Particles, Sources and Fields*, vol. 2. Addison-Wesley
27. P.V. Landshoff, J.C. Polkinghorne, J.C. Taylor, *A Proof of the Mandelstam Representation in Perturbation Theory*, Nuovo Cimento, vol. 19, p. 939 (1961)
28. R.J. Eden, Proof of Mandelstam representation for every order in perturbation theory. Phys. Rev. **121**, 1567 (1961)
29. R.J. Eden, P.V. Landshoff, J.C. Polkinghorne, J.C. Taylor, Acnodes and cusps on landau curves. J. Math. Phys. **2**, 656 (1961)
30. R.J. Eden, P.V. Landshoff, D.I. Olive, J.C. Polkinghorne, *The Analytic S-matrix* (Cambridge University Press, 1966)

Some Uses and Applications of Analyticity and Dispersion Relations

12

12.1 Low Energy Meson-Nucleon Scattering

Among the earliest applications of Dispersion Relations were that of Chew, Goldberger, Low and Nambu [1] to the problem of low energy meson-nucleon scattering. Meson-nucleon interactions were the ones that posed serious challenges to QFT's, and it was the hope of the analytic S-matrix programme to succeed where QFT had failed.

As with many papers on dispersion relations, there is a large amount of algebra which is often context-specific and not very illuminating, though essential. We shall skip all such details and only highlight the essential aspects. In a sense, the pion-nucleon scattering amplitude is among the most complicated as it involves both spin as well as isospin. Denoting the nucleon momenta by p_1 (initial) and p_2 (final), and likewise the pion momenta by q_1, q_2 , the independent four vectors are chosen to be $P = \frac{1}{2}(p_1 + p_2)$, $Q = \frac{1}{2}(q_1 + q_2)$, $\kappa = \frac{1}{2}(q_1 - q_2)$, while the two independent invariants to be $\nu = -\frac{P \cdot Q}{M}$, κ^2 . This is before the Mandelstam variables had become popular. The T -matrix (also a matrix in spinor as well as charge space) is decomposed as (with γ being the gamma matrices)

$$T = -A + i \gamma \cdot Q B \quad (12.1)$$

A and B are functions of the two invariants. If α, β are the isospin of the initial and final mesons, there is a T for each pair $\alpha\beta$, and consequently $A_{\beta\alpha}, B_{\beta\alpha}$ also, each of which is a 2×2 matrix. To bring out the independent invariants in charge (read isospin) space, they further introduce

$$A_{\beta\alpha} = \delta_{\beta\alpha} A^{(+)} + \frac{1}{2} [\tau_\beta, \tau_\alpha] A^{(-)} \quad B_{\beta\alpha} = \delta_{\beta\alpha} B^{(+)} + \frac{1}{2} [\tau_\beta, \tau_\alpha] B^{(-)} \quad (12.2)$$

Thus there are four independent invariant functions characterizing the scattering, i.e. $A^{(\pm)}$, $B^{(\pm)}$. Their significance can be further elucidated by decomposing the channels to those with total isotopic spins, which in this case are $I = \frac{3}{2}$ and $I = \frac{1}{2}$.

$$A^{(+)} = \frac{1}{3} (A^{(\frac{1}{2})} + 2 A^{(\frac{3}{2})}) \quad A^{(-)} = \frac{1}{3} (A^{(\frac{1}{2})} - A^{(\frac{3}{2})}) \quad (12.3)$$

and likewise for $B^{(\pm)}$. Thus the four invariants correspond to no spin-flip, spin-flip in each of the two isospin channels.

An excellent account of these and other details can also be found in Chap. 23 of [2], where dispersion relations for forward pion-nucleon scattering are discussed at length. All the invariant amplitudes are functions of ν , κ^2 . Of these, ν plays the role of energy of scattering, and κ^2 the momentum transfer. The authors then set up dispersion relations for each of them. Specifically, they set up these relations for the real parts in terms of integrals over their imaginary parts which, through their relations to total cross sections, are in principle observable. In order to make progress, they make several simplifying assumptions. The most drastic of them is that the invariant amplitudes obey *unsubtracted* dispersion relations. They further assume the dominance of the $(3, 3)$ (that is, isospin $\frac{3}{2}$ and spin $\frac{3}{2}$) resonance. They further restrict the range of the integrals to the vicinity of this resonance.

After a partial-wave analysis, they obtain phase shifts for the S, P, D-wave channels in terms of a small number of parameters. For further details, the reader is recommended to follow their paper [1].

12.2 Pion Decay

The next important application of dispersion relations that we consider is the work of Goldberger and Treiman to the decay of charged pions [3]. The results of this analysis had one of the most important and long-lasting impacts of dispersion relations in the form of the celebrated *Goldberger-Treiman Relation* which became the spring board for many path-breaking developments like chiral symmetry, partially conserved axial currents (PCAC), etc. These developments are discussed at length in the coming chapters.

In their work, they study the charged pion decays $\pi^\pm \rightarrow l^\pm + \nu$ where l is either an electron or a muon. From the earlier analysis of this decay by Ruderman and Finkelstein [4] as well as observationally, the charged pions decay predominantly into muons rather than into the much lighter electrons. This was one of the strongest evidences for the role of axial vector currents in this decay, hence the importance of this decay in the development of chiral symmetry, PCAC, etc. For this reason, Goldberger and Treiman only considered the decays $\pi^\pm \rightarrow \mu^\pm + \nu$ (we have clubbed both neutrinos and anti-neutrinos into a single ν).

They treat the pions as “in” and the $\mu\nu$ pair as “out” states in the spirit of the Lehmann-Symanzik-Zimmermann (LSZ) formalism which has been extensively developed in Chap. 8 of our book. On “reducing” the neutrino field, they obtain

$$\langle \mu\nu, out | \pi, in \rangle = i(2\pi)^4 \delta^{(4)}(p_\mu + p_\nu - p_\pi) \langle \mu | \bar{f}_\nu(0) | \pi \rangle (1 + \gamma_5) u(p_\nu) \quad (12.4)$$

They next write out the coefficient of $(1 + \gamma_5)u(p_\nu)$, after reducing the pion field also, as

$$\begin{aligned} \mathcal{M} &\equiv \sqrt{\frac{E_\mu}{m_\mu}} \sqrt{2E_\pi} \langle \mu | \bar{f}_\nu | \pi \rangle \\ &= i \sqrt{\frac{E_\mu}{m_\mu}} \int dx e^{-i(p_\pi - p_\mu) \cdot x} \langle \mu | T(\bar{f}_\nu(x) J(0)) | 0 \rangle \end{aligned} \quad (12.5)$$

This is the T -product representation discussed in Chap. 8. They also obtain the retarded-product representation for the physical decay amplitude (also discussed in Chap. 8):

$$\mathcal{M} = i \sqrt{\frac{E_\mu}{m_\mu}} \int dx e^{-ip_\nu \cdot x} \theta(x) \langle \mu | [\bar{f}_\nu(x), J(0)] | 0 \rangle \quad (12.6)$$

where J is the source for the pion field (after making use of the overall energy-momentum conservation). We have just reproduced the reasonings of Goldberger and Treiman here. The reader is urged to check the details in the light of the discussions in Chap. 8 on LSZ reduction formulae. From Lorentz invariance it follows that this must be of the form

$$\mathcal{M} = F(p_\pi^2) \bar{u}(p_\mu) \gamma_5 \gamma_\alpha p_\pi^\alpha \quad (12.7)$$

They then show that F satisfies a dispersion relation. They provide a justification in the appendix to their paper. The form of the dispersion relation they propose is

$$F(\xi) = \frac{1}{\pi} \int_{-\infty}^{\infty} d\xi' \frac{\text{Im } F(-\xi')}{\xi' + \xi - i\epsilon} \quad (12.8)$$

They arrive at their form of dispersion relations on the basis of a very specific form for the neutrino source which is equivalent to a purely axial vector current coupling between leptons and nucleons. Since the vector current does not contribute to pseudoscalar pion decay, this may not be in conflict with the $V - A$ theory. They further claim that with this specific form of coupling between leptons and nuclei, the dispersion relations are unsubtracted. It may be possible to arrive at their form of dispersion relations on the basis of more general considerations like the Jost-Lehmann-Dyson theorem discussed in Chap. 10.

They arrive at the absorptive part of F by the general techniques already discussed in Chap. 8, i.e. by introducing a complete set of sets in the commutator. They make

drastic assumptions about how this sum over intermediate states is saturated. If parity is conserved, the lowest mass intermediate state that contributes is the three-pion state. But they assume that it is the considerably more massive nucleon-antinucleon states that dominate. For this they do not produce very convincing arguments. Part of their justification is that the contribution of this channel can be estimated in terms of observables when the lepton-nucleon coupling has been assumed as before.

Then follows a lot of algebra whose discussion we omit. Their punch-line is their result for $F(0)$ which is presumed to be a good approximation for the physically relevant $F(-m_\pi^2)$ by virtue of the small pion mass:

$$F(0) = -\frac{m}{2\pi^2} \sqrt{2} G g_A \frac{J}{1 + \left(\frac{G^2}{4\pi}\right) \left(\frac{2J}{\pi}\right)} \quad (12.9)$$

where G is the pion-nucleon coupling constant with the experimentally determined value $\frac{G^2}{4\pi} \approx 15$, $g_A = \frac{G_A}{G_V}$, and m the nucleon mass. As it stands, $F(0)$ depends on J , a quantity dependent on dynamics. But for large values of J , it becomes J -independent! J does not have to be very large either; it suffices for $J \gg 0.1$ to reach the asymptotic value of

$$F(0) = -\sqrt{2} m \frac{g_A}{G} \quad (12.10)$$

$F(0)$ is essentially the so-called *pion decay constant* F_π modulo some normalization factors. This famous relation has stood the test of time and is also very well satisfied phenomenologically, notwithstanding various lacunae in its treatment. As we shall see later, there are better ways of arriving at this relation (as discussed in Chap. 18) which also clarify the meaning and significance of this relation. The present derivation was also criticized by Gell-Mann and Levy in their classic paper [5] as also by Sawyer [6]. Both Gell-Mann and Levy [5], as well as Nambu [7], gave a new meaning and perspective to the Goldberger-Treiman relation that led to the birth of the concept of partially conserved axial currents (PCAC) and eventually to the central theme of chiral symmetry in strong interaction physics.

12.3 The Froissart, Khuri-Kinoshita Bounds

This section draws its inspiration from Andre Martin's lectures on *The Rigorous Analyticity-Unitarity Programme and its Successes* at the Ringberg Symposium on QFT in honour of Wolfhart Zimmermann, June 1998 [8]. Among the most impressive of these successes is the *Froissart Bound* on scattering amplitudes and cross sections [9]. Froissart's analysis is based on the Mandelstam Double Spectral representation (see Chap. 11) and the structure of its possible subtractions. Based on certain parallels between the Mandelstam Representation and Potential Scattering, Froissart proves certain features of the high energy behaviour of amplitudes and scattering cross sections.

Summarizing these results (in Froissart's own words): "At forward or backward angles, i.e. $|z| = |\cos \theta| = 1$, the modulus of the amplitude behaves at most like $s \ln s^2$, as s goes to infinity. We can use the optical theorem to derive that total cross sections behave at most like $\ln^2 s$. At non-forward angles, the amplitude behaves at most like $s^{\frac{3}{4}} \ln^{\frac{3}{2}}(s)$ ". Khuri, in his Brookhaven Summer School lectures [10] gives an expression that makes the dependence on the scattering angle explicit:

$$|A(s, \cos \theta)| \leq \frac{s \ln^{\frac{3}{2}}(s)}{\sin^{\frac{1}{2}} \theta} \quad (12.11)$$

Martin narrates, in the same lectures, his own approaches to this problem [11, 12] which does not use the Mandelstam Representation, but relies instead on enlarging both the small and large Lehmann ellipses (see Chap. 10) which eventually lead to the more accurate bound [13]:

$$\sigma_T < \frac{\pi}{m_\pi^2} (\ln s)^2 \quad (12.12)$$

In particular, he invokes what may be called *Martin Analyticity*: with s fixed *near* a point in the physical region, assuming fixed- t dispersion relations, positivity as well as *BEG analyticity* [14], the amplitude $F(s, t)$ is analytic in $|t| < R$, with R being *independent* of s . A crucial intermediate result was that for $|t| < 4m_\pi^2$, the number of subtractions in the dispersion relations was at most two [15].

It is important to appreciate that what both Froissart and Martin proved were *asymptotic bounds*, and not *asymptotic estimates*. The actual behaviour of amplitudes and cross sections at high energies could still be something quite different, as long as it is consistent with these bounds. They could be rising, constant, or, even falling! It is indeed incredible that reasonings based on analyticity could so severely constrain something that is so manifestly *Dynamic*!

Khuri and Kinoshita [16] showed that if total cross sections indeed saturated the Froissart bound, i.e. if $\sigma_T \simeq \ln^2 s$, the following would be true too:

$$\frac{\text{Re} F(s)}{\text{Im} F(s)} \simeq \frac{\pi}{\ln s} \quad (12.13)$$

12.3.1 The Pomeranchuk Theorem

Another remarkable application of analyticity is the so-called *Pomeranchuk Theorem*. This theorem concerns the asymptotic equality for the total cross sections $\sigma_T(AB)$ and $\sigma_T(A\bar{B})$, where $\bar{B}$ is the antiparticle of B . That analyticity principles can constrain the behaviour of antiparticles should not be totally surprising as crossing symmetry binds these two together.

In his original work Pomeranchuk [17] had analysed the situation under the assumption that total cross sections remain finite. With some additional assumptions about the real parts of scattering amplitudes, he could show that

$$\sigma_T(AB) \rightarrow \sigma_T(A\bar{B}) \quad s \rightarrow \infty \quad (12.14)$$

Eden [18], and, Kinoshita [19], reanalyzed this problem under the more realistic rising total cross sections. They were able to show that

$$\frac{\sigma_T(AB)}{\sigma_T(A\bar{B})} \rightarrow 1 \quad s \rightarrow \infty \quad (12.15)$$

12.3.2 Pi-Pi Scattering

As another example of the power of analyticity, Martin cites the example of Pion-Pion scattering. Employing his ideas of extended analyticity (meaning beyond those of the Lehmann Ellipses) along with *playing with unitarity and crossing in a clever way*, he gets a bound on the scattering amplitude at the *symmetry point* $s = t = u = \frac{4m_\pi^2}{3}$:

$$|F(s = t = u = \frac{4m_\pi^2}{3})| < 4 \quad (12.16)$$

The amplitude above is normalized such that $F(s = 4m_\pi^2, t = 0, u = 0)$ is the scattering length a_{00} .

12.3.3 Some Recent Developments

Already in his Ringberg lectures of 1998 Martin mentions increasing experimental support to all these striking results. But today the energies at which, say, proton-proton or proton-antiproton collisions are carried out, are phenomenal. There is also renewed interest in the above-mentioned topics particularly from the point of view reconciling them with a theory like *Quantum Chromodynamics* (QCD) where the asymptotic particles, which are color-neutral hadrons, do not bear straightforward relations to the underlying quantum fields of Quarks and Gluons, in view of the difficulties in a controlled treatment of *Color Confinement*.

One such interesting revival is a relatively recent paper by Veronica Errasti Diez, Rohini Godbole and Aninda Sinha [20]. They analyse the Froissart bounds as well as Pomeranchuk theorem from the point of view of *Gauge-Gravity Duality*. In particular, they find a correction of the type $\ln \frac{s}{s_0}$ with a *negative coefficient* to the leading order Froissart bound. A nice feature of their paper is a detailed analysis of the experimental situation that includes LHC data at $\sqrt{s} \simeq 8$ -TeV as well as high energy cosmic ray data. They find agreement with both the magnitude and sign of the correction terms. They also compare the data with predictions of the (improved) Pomeranchuk

Theorem. They mention the interesting fact that the difference $\sigma_T(\bar{p}p) - \sigma_T(pp)$ has a $s^{-\frac{1}{2}}$ behaviour [21].

While it is interesting to look at this from Gauge-Gravity Duality, AdS-CFT correspondence, etc. (we shall see later that such considerations have also been invoked to understand QCD Strings) it is imperative to understand them as directly from QCD as possible. The authors of [20] also emphasize the desirability of this, but think that a non-perturbative analysis of QCD may be necessary which is undoubtedly very hard. But since only the asymptotically very high energy scatterings need to be understood, the asymptotic freedom of QCD whereby coupling constants decrease at high energies a perturbative QCD calculation may suffice.

In another of very recent revivals of the topic of Froissart bounds, Halder and Sinha [22] consider obtaining Froissart Bounds from Mellin Transforms in Conformal Field Theories. All these revivals are indeed fascinating and it will be interesting to see if they can lead to the much sought after proof of the Mandelstam Double Spectral Representation.

12.4 Adler-Weisberger Relations for g_A

We end this chapter with a discussion of a rather remarkable application of dispersion relations by Adler [23] and Weisberger [24], who actually calculated the renormalized value of g_A in terms of pion-nucleon cross sections, albeit at off-shell values. But they find good ways of approximating the off-shell quantities by measurable on-shell equivalents and find numerical values for g_A in very good agreement with data. Adler and Weisberger obtained the values 1.24 and 1.15, respectively. At that time, the neutron lifetime, which determines g_A was about 11.8 min, and the inferred value of g_A was 1.18. But the neutron lifetime estimates have changed radically since then! The currently accepted value is 880s with an unexplained discrepancy of 8s between free and bottled neutrons, leading to an estimate of $g_A \approx 1.2756$ [25]. Needless to say, the value of g_A impacts the accuracy of the Goldberger-Treiman relation also. The importance and relevance of these issues are again taken up in section Sect. 18.4 of Chap. 18.

Now we present the salient aspects of the derivations by Adler and Weisberger. Both of them are meticulous and have paid attention to many important details. For example, Weisberger explicitly comments on the more singular terms present in equal time commutators and the consequent appearance of derivatives of delta functions. We have also commented on such issues at various parts of the book. Weisberger argues that in the limits relevant to the calculations, they do not contribute. We shall not go into such details, nor into the details of the numerical estimates. Both the works are based on the following ingredients: (i) the nature of the hadronic currents, (ii) Gell-Mann's current algebra [26], and, (iii) the partially conserved axial current(PCAC) hypothesis. The hadronic currents of importance, according to the $V - A$ theory, are the vector currents V_μ^a and the axial vector currents A_μ^a . They can be further classified into $\Delta S = 0$ and $\Delta S \neq 0$ parts where S is the strangeness. At a time when the other quarks like charm, bottom and top were not known, the $\Delta S = 0$ part

of the currents had the weight factor of $\cos \theta$ with θ being the Cabibbo angle. In the current picture, the Cabibbo angle gets generalized to the Cabibbo-Kobayashi-Maskawa (CKM) unitary matrix. Weisberger has analysed the $\Delta S = 1$ contributions also. The two works are complementary to each other. In this section, we shall only illustrate Adler's approach for the $\Delta S = 0$ part. Adler's expression for the hadronic currents erroneously includes the Fermi constant $G_F = G_V$, but this does not affect the correctness of his calculations.

As clarified later in Chap. 18, the explicit forms of the vector and axial vector currents in terms of nucleon and pion fields are not necessary as only their matrix elements between nucleon states will be used and they can be parametrized using the invariances of the theory in terms of G_V and G_A (we have rewritten Adler's expression in a form symmetric between initial and final momenta):

$$\langle N(q) | A_\mu^+(x) | N(q') \rangle = g_A \sqrt{\frac{M_N^2}{q_0 q'_0}} \bar{u}_N(q) \gamma_\mu \gamma_5 \tau^+ u_N(q') \quad (12.17)$$

The essential aspect of Gell-Mann's current algebra, which is crucial to the derivations, is the equal time commutation relation (ETCR) for the axial vector current

$$[A_4^a(x), A_4^b(y)]|_{x_0=y_0} = -\delta^{(3)}(\mathbf{x} - \mathbf{y}) \epsilon^{abc} V_4^c(x) \quad (12.18)$$

where A_μ^a, V_μ^a are the axial and vector currents respectively. As remarked by Adler, the precise form of this ETCR is not really required as long as the integrated form of it for the axial charges (called chiralities by Adler)

$$[\chi^+(t), \chi^-(t)] = 2 I_3 \quad (12.19)$$

holds. The chiralities are defined by (in Adler's choice of metric)

$$\chi^\pm(t) = -i \int d^3x (A_4^1 \pm i A_4^2) \quad (12.20)$$

The vector currents V_μ^a are conserved by the CVC hypothesis and hence Isospin I_3 is conserved. On the other hand, axial currents are only partially conserved as per PCAC:

$$\partial_\mu A_\mu^a = \frac{M_N M_\pi^2 g_A}{g_r K_{NN\pi}(0)} \phi^a(x) \quad (12.21)$$

(in the notation of Adler). Here g_r is the renormalized pion-nucleon coupling constant with the approximate phenomenological value of $\frac{g_r^2}{4\pi} \approx 14.6$, $K_{NN\pi}$ is the pion-nucleon form factor normalized such that $K_{NN\pi}(-M_\pi^2) = 1$, ϕ^a the iso-triplet pion field, and, M_N, M_π the nucleon and pion masses, respectively. So for all practical purposes $K_{NN\pi}(0)$ can be taken to be unity. Such an extrapolation was used by Goldberger and Treiman also in [3]. It is customary to express this in terms of the pion decay constant f_π as

$$\partial_\mu A_\mu^a = f_\pi M_\pi^2 \phi^a \quad (12.22)$$

The two forms can be reconciled upon using the Goldberger-Treiman relation. This PCAC relation can be integrated to give the time-dependence of $\chi^\pm(t)$:

$$\frac{d}{dt} \chi^\pm(t) = \frac{\sqrt{2} M_N M_\pi^2 g_A}{g_r K_{NN\pi}(0)} \int \phi^\pm(x) \quad (12.23)$$

In quantum theory the χ 's are operators and this should be thought of as $\frac{i}{\hbar} [H, \chi^\pm(t)]$ with H the total Hamiltonian.

The next step is to take the matrix elements of Eq. (12.19) between single proton states and evaluate the l.h.s with the help of the equations given above. We shall only illustrate this so-called *Fubini-Furlan Method*. Adler discusses yet another derivation of his sum rule and Weisberger's approach provides a third method. The r.h.s is given by

$$\langle p(q) | 2 I_3 | p(q') \rangle = (2\pi)^3 \delta^{(3)}(\mathbf{q} - \mathbf{q}') \quad (12.24)$$

(in this confusing notation of Adler $|p(q)\rangle$ is a proton state with momentum $\mathbf{q}$!). The l.h.s is evaluated by the standard trick of introducing a complete set of states. Isolating the contribution of the single neutron states this is written as

$$\begin{aligned} \langle p(q) | [\chi^+(t), \chi^-(t)] | p(q') \rangle &= \sum_{spin} \int \frac{d^3k}{(2\pi)^3} \langle p(q) | \chi^+(t) | n(k) \rangle \langle n(k) | \chi^-(t) | p(q') \rangle \\ &+ \sum_{j \neq N} \langle p(q) | \chi^+(t) | j \rangle \langle j | \chi^-(t) | p(q') \rangle - (\chi^+ \leftrightarrow \chi^-) \end{aligned} \quad (12.25)$$

The contribution from the neutron intermediate states, as calculated from using Eq. (12.17), is (we refer the reader to Adler's paper for details)

$$(2\pi)^3 \delta^{(3)}(\mathbf{q} - \mathbf{q}') g_A^2 \left(1 - \frac{M_N^2}{q_0^2} \right) \quad (12.26)$$

The rest of the contributions, on making use of Eq. (12.23) (in the $\frac{i}{\hbar} [H, \chi^\pm(t)]$ form), can be recast as

$$\frac{2 M_N^2 M_\pi^4 g_A^2}{g_r^2 K_{NN\pi}^2(0)} \sum_{j \neq N} \frac{\langle p(q) | \int d^3x \phi^+(x) | j \rangle \langle j | \int d^3x \phi^-(x) | p(q') \rangle}{(q_0 - q_{j0})^2} - (\pi^+ \leftrightarrow \pi^-) \quad (12.27)$$

As an intermediate step (it is not necessary to invoke this, though it simplifies the calculations) Adler uses

$$\sum_{j \neq N} = \sum_{j \neq N}^{\prime} \int_{M_N + M_\pi}^{\infty} dW \delta(W - M_j) \int \frac{d^3q_j}{(2\pi)^3} \quad (12.28)$$

where $\prime$ denotes summation(integration) over the internal degrees of freedom of the intermediate states j . The expression given by Adler in his Eq. (12) has been rewritten to make it more sensible. The matrix elements of the pion fields are written as

$$\langle j | \phi^\pm(0) | p(q) \rangle = \sqrt{\frac{M_N M_j}{q_0 q_{j0}}} F_j^\pm \quad (12.29)$$

At this point, it is worth drawing attention to a remarkable feature of these *sum rules*. This is already manifest in Eq. (12.19) itself; while the l.h.s depends on time, the r.h.s is time-independent. In other words, there is an infinity of these sum rules for each t . This translates into an infinity of sum rules for each value of q_0 . There are various momentum conservation delta functions arising out of the x -integrations above and the net result is equating $\mathbf{q}_j$ to $\mathbf{q}$. On using this it follows that $q_0^2 - M_N^2 = q_{j0}^2 - M_j^2$ which translates into two important relations

$$q_{j0} = \sqrt{(q_0^2 + M_j^2 - M_N^2)} \quad (q_0 - q_{j0})^{-2} = \frac{(q_0 + q_{j0})^2}{(M_j^2 - M_N^2)^2} \quad (12.30)$$

Next, Adler chooses to take the $q_0 \rightarrow \infty$ limit of the sum rules. This is of course fine as long as the limit is well defined, but it nevertheless leaves unanswered the meaning and consistency of the sum rules for all other values of q_0 . Adler also had to use this limit interchangeably with various summations and integrations. He gives a second derivation of the sum rules to justify such potentially dangerous steps. On defining

$$K^\pm(W, (q - q_j)^2) = \sum_{j \neq N}^{\prime} \delta(W - M_j) M_\pi^4 |F_j^\pm|^2 \quad (12.31)$$

and taking the $q_0 \rightarrow \infty$ limit, Adler arrives at the intermediate sum rule:

$$1 - \frac{1}{g_A^2} = \frac{2M_N^2}{g_r^2 K_{NN\pi}(0)^2} \int_{M_N+M_\pi}^{\infty} dW \frac{4M_N W}{(W^2 - M_N^2)^2} [K^+(W, 0) - K^-(W, 0)] \quad (12.32)$$

Of course, the $K^\pm(W, 0)$ are not by themselves observable. In the last step of the derivation, Adler relates these to pion-proton scattering (albeit off-shell). We leave out the details and cite his final result:

$$\sigma_0^\pm(W) = \frac{2\pi M_N}{(W^2 - M_N^2)} K^\pm(W, 0) \quad (12.33)$$

The $\sigma_0^\pm(W)$ are total cross sections for $\pi^\pm$ -proton scattering at centre of mass energy W . But as they stand, these are *off massshell* and not directly observable. The final form of the sum rule becomes

$$1 - \frac{1}{g_A^2} = \frac{4M_N^2}{g_r^2 K_{NN\pi}(0)^2} \frac{1}{\pi} \int_{M_N+M_\pi}^{\infty} \frac{W dW}{W^2 - M_N^2} [\sigma_0^+(W) - \sigma_0^-(W)] \quad (12.34)$$

A very remarkable result indeed. Adler then goes on to numerically evaluating the r.h.s of the sum rule (in section III of his paper) and shows how it can be split into three terms, the first of which can be estimated from physical on-shell scattering cross sections, the second of which arises from finite pion mass, and lastly, a third term which is model dependent. Fortunately, the last term is estimated to be only 15% of the total. The value of g_A arrived at by Adler has already been mentioned.

References

1. G.F. Chew, M.L. Goldberger, F.E. Low, Y. Nambu, Phys. Rev. **106**, 1377 (1957)
2. S. Gasiorowicz, *Elementary Particle Physics* (Wiley)
3. M.L. Goldberger, S.B. Treiman, Phys. Rev. **110**, 1178 (1958)
4. M.A. Ruderman, R. Finkelstein, Phys. Rev. **76**, 1458 (1948)
5. M. Gell-Mann, M. Levy, Nuovo Cimento **16**, 705 (1960)
6. R.F. Sawyer, Phys. Rev. **116**, 231 (1959)
7. Y. Nambu, Phys. Rev. Lett. **4**, 380 (1960)
8. A. Martin, *Ringberg Symposium on Quantum Field Theory* (1998)
9. M. Froissart, Phys. Rev. **123**, 1053 (1961)
10. N.N. Khuri, *Summer School in Elementary Particle Physics* (Brookhaven National Lab., 1969)
11. A. Martin, Phys. Rev. **129**, 1432 (1963)
12. A. Martin, Nuovo Cimento **42**, 901 (1966); Nuovo Cimento **42**, 930 (1966)
13. Y. Lukaszuk, A. Martin, Nuovo Cimento **52**, 122 (1967)
14. J. Bros, H. Epstein, V. Glaser, Comm. Math. Phys. **1**, 240 (1965)
15. Y.S. Jin, A. Martin, Phys. Rev. B **135**, 1375 (1964)
16. N.N. Khuri, T. Kinoshita, Phys. Rev. B **137**, 720 (1965)
17. I.Ya. Pomeranchuk, Sov. Phys. JETP **7**, 499 (1958)
18. R.J. Eden, Phys. Rev. Lett. **16**, 39 (1966)
19. T. Kinoshita, *Perspectives in Modern Physics*, in ed. by R.E. Marshak (New York, 1966), p. 211
20. V. Errasti Diez, R. Godbole, A. Sinha, Phys. Lett. B **746**, 285 (2015)
21. M.M. Block, R.N. Cahn, Phys. Lett. B **126**, 224 (1983)
22. P. Halder, A. Sinha, SciPost Phys. **8**, 093 (2020)
23. S.L. Adler, Phys. Rev. **140**, B736 (1965)
24. W.I. Weisberger, Phys. Rev. **143**, 1302 (1966)
25. W.J. Marciano, A. Sirlin, (2018). [arXiv:1802.01804](https://arxiv.org/abs/1802.01804) [hep-th]
26. M. Gell-Mann, Physics **1**, 63 (1964)



In the Land of Complex Angular Momentum

13

13.1 Introduction

We saw in Chap. 10 how scattering amplitudes could be analytically continued in scattering angle, or equivalently in the Mandelstam variable t away from the physical region. In particular, we saw how the domain of analyticity in $z = \cos \theta$ could be extended to the small Lehmann Ellipse for the scattering amplitude, and the large Lehmann Ellipse for its imaginary part $\text{Im } T$.

In this chapter we explore some of the deeper implications of these Lehmann Analyticities to the well-known *partial-wave expansions*. In fact, partial-wave analysis brings out some deeper aspects of the Lehmann Ellipses. In particular, the Lehmann Ellipses dictate the large- l behaviour of partial-wave amplitudes in a well-defined way.

In a remarkable extension of Lehmann Analyticity, it turns out to be possible to find a representation of the scattering amplitude which is regular in a domain that is significantly larger than the Lehmann Ellipses. The price one has to pay for this is an extension *nonphysical*, even complex, values of the angular momentum l . Though this might appear to be a purely mathematical device without any obvious physical significance, the notion of singularities in the complex- l plane, in particular the so-called *Regge Poles*, named after Tulio Regge who discovered them in the context of *Potential Scattering* [1], control not only the low energy resonances in the direct channel, but also the high energy behaviour in the cross-channels. It is this unifying feature of Regge Poles that led to the concept of *Duality* that forms the back-bone of the topics elaborated in this book.

As noted by Peter Goddard in his reminiscences of Stanley Mandelstam [2], the book *The analytic S-matrix* [3] begins with the rather hyperbolic *One of the most important discoveries in elementary particle physics has been that of the complex plane*. But close to six decades later, the impressive developments have borne testimony to the essential correctness of this!

13.2 Lehmann Ellipses and Partial-wave Analysis

Following Lehmann [4], we start with the partial-wave expansion for the elastic scattering of two spin-less particles:

$$T(W, \cos \theta) = \frac{1}{\pi^2} \frac{W}{K} \sum_{l=0}^{\infty} (2l+1) C_l(W) P_l(\cos \theta) \quad (13.1)$$

The inverse is given by

$$C_l(W) = \frac{\pi^2}{2} \frac{K}{W} \int_{-1}^1 dz T(W, z) P_l(z) \quad (13.2)$$

Here W and K are respectively the centre of mass energy and momentum, and, $z = \cos \theta$. Lehmann uses a normalization that is different from what may be found in many treatments of the partial-wave expansions. Those can be obtained on using $a_l(W) = \frac{W}{K} \frac{C_l(W)}{\pi^2}$.

When the scattering amplitude is analytically extended to the Lehmann Ellipse (let us consider the small Ellipse first for both real and imaginary parts of T), the argument z of the Legendre polynomials is in general complex, with a magnitude that could exceed unity. It is known from the theory of Legendre functions that the large- l asymptotics can then be dramatically different. The large- l asymptotics (for real and positive l) in question is given by [5, p.162]:

$$|P_l(z)| < l^{-1/2} e^{l |\operatorname{Im} \theta|} f(z) \quad l \rightarrow \infty \quad (13.3)$$

Thus even a tiny imaginary part to θ can cause an *exponential* increase! As a consequence, for the partial-wave expansion to converge for complex angles, the partial-wave amplitudes $C_l(W)$ (or $a_l(s)$) must decay exponentially for large- l . Let us suppose that $C_l(W) \simeq e^{-\Gamma l}$ for very large l (real and positive). Then it follows that $\Gamma > \operatorname{Im} \theta$. But Γ being a property of the partial-wave amplitudes can only depend on the Mandelstam invariant s . It is at this stage that the Lehmann Ellipses play a crucial role.

Recall our earlier discussion about how ellipses occur generically when angles are extended to complex values. If $\theta = \theta_r + i \theta_i$,

$$z = \cos \theta = \cos(\theta_r + i \theta_i) = \cos \theta_r \cosh \theta_i - i \sin \theta_r \sinh \theta_i \quad (13.4)$$

Hence

$$\left(\frac{\operatorname{Re} z}{\cosh \theta_i} \right)^2 + \left(\frac{\operatorname{Im} z}{\sinh \theta_i} \right)^2 = 1 \quad (13.5)$$

showing that in the complex- z plane one gets an ellipse with semi-major axis $\cosh \theta_i$, semi-minor axis $\sinh \theta_i$. Consequently, the foci are at ± 1 . The sum of the two axes,

$\cosh \theta_i + \sinh \theta_i = e^{\theta_i}$, yields θ_i . For the small Lehmann ellipse, the semi-major axis $x_0(s)$ is given by

$$x_0(s) = \left\{ 1 + \frac{(m_1^2 - \mu^2)(m_2^2 - M^2)}{K^2[s - (m_1 - m_2)^2]} \right\}^{\frac{1}{2}} \quad (13.6)$$

We remind the reader that s is the square of the centre of mass energy, K the centre of mass momentum, μ , M the masses of Mesons and Nucleons respectively, m_1 , m_2 the lowest masses of multiparticle states with the quantum numbers of a single meson and nucleon plus mesons.

It immediately follows that Γ above is given by

$$\Gamma_s = \text{Im } \theta_s = \ln(x_0(s) + \sqrt{x_0^2(s) - 1}) \quad (13.7)$$

leading to the very important conclusion (see Eq. (13a) of [4]):

$$|\text{Re } C_l(W)| \simeq (x_0(s) + \sqrt{x_0^2(s) - 1})^{-l} \text{ as } l \rightarrow \infty \quad (13.8)$$

It is to be appreciated that Lehmann Ellipses, which Lehmann established solely on the basis of the LSZ formalism and the *Integral Representations* obtained by the Jost-Lehmann [6] and Dyson [7], are essential for obtaining this important result.

Now we turn to an analogous treatment of the partial-wave expansions for $\text{Im } T(W, \cos \theta)$. Before proceeding, an important caveat ought to be mentioned. This concerns the order in which the scattering angle is complexified and the imaginary part calculated. It would be incorrect to first complexify scattering angles in Eq. (13.1) and then take the imaginary part of T ; the correct procedure is to first take the imaginary part of this equation for real θ to get

$$\text{Im } T(W, \cos \theta) = \frac{1}{\pi^2} \frac{W}{K} \sum_{l=0}^{\infty} (2l+1) \text{Im } C_l(W) P_l(\cos \theta) \quad (13.9)$$

and *then* complexify θ . After that, the same line of reasoning as above can be adopted to arrive at the large- l asymptotics of $\text{Im } C_l(W)$ by using Lehmann's results for the Large Lehmann Ellipses instead. For the case of elastic scattering, the semi-major and semi-minor axes of the large Lehmann Ellipse are given by $y_0(s) = 2x_0^2(s) - 1$ and $\sqrt{y_0^2(s) - 1} = 2x_0(s)\sqrt{x_0^2(s) - 1}$, respectively. On noting that $y_0(s) + \sqrt{y_0^2(s) - 1} = (x_0(s) + \sqrt{x_0^2(s) - 1})^2$,

$$\Gamma_L = \text{Im } \theta_L = \ln(y_0(s) + \sqrt{y_0^2(s) - 1}) = 2\Gamma_s \quad (13.10)$$

and finally,

$$|\text{Im } C_l(W)| \simeq (x_0(s) + \sqrt{x_0^2(s) - 1})^{-2l} \text{ as } l \rightarrow \infty \quad (13.11)$$

which is Eq. (14a) of [4].

13.3 Going Beyond Lehmann Ellipses: Complex Angular Momentum

As should have become clear now, the quest is for higher and higher degrees of analyticity, or equivalently, larger and larger domains of analyticity. The immediate need is to go beyond the Lehmann Ellipse analyticities. Most amazingly, the path to this turns out to be an analytic continuation of the partial-wave amplitudes $C_l(W)$ to *complex* values of the angular momentum l . Whether such a continuation is at all possible, and if so how to carry out the continuation is technically rather complicated, and it occupied sustained research in the early days. There are several excellent sources available for the interested reader; to cite a few, *The Analytic S-matrix* [3], all the lectures of the *The Brookhaven Summer School in Elementary Particle Physics, Brookhaven National Lab., July 22-Aug 29, 1969* [8], *Dual Resonance Models* by P.H. Frampton [9], and, *Complex Angular Momenta* by E.J. Squires [10]. We shall mostly follow Squires book as it has an especially lucid and succinct account. For our purposes, essentially Chap. 1 is all we need. It also has an excellent reprinting of some of the central papers.

We shall follow Squires in analysing the partial-wave amplitude $a_l(s)$ related to Lehmann's $C_l(W)$. The idea is to find a function $a(l, s)$ where l now is a complex variable such that $a(l, s) = a_l(s)$ for positive semidefinite integral values of l . The central question is whether the knowledge $a_l(s)$ for all the positive integral values of l can determine $a(l, s)$ uniquely. The answer is in the negative in general. It turns out, however, that by specifying the asymptotic behaviour of $a(l, s)$ for $l \rightarrow \infty$, the continuation can be made unique.

This is made possible by *Carlson's Theorem* which is stated as follows: Given a $f(z)$, *holomorphic* in $\text{Re } z \geq 0$, such that $f(z) = 0$ for $z = 0, 1, \dots$, and $f(z) = O(e^{\lambda|z|})$ for $|z| \rightarrow \infty$ in $\text{Re } z \geq 0$, with $\lambda < \pi$, then $f(z) = 0$ in $\text{Re } z \geq 0$.

A word of caution that z in the statement of the theorem really should be thought of as l , and not z as in $P_l(z)$! Squires refers to Sect. 5.81 of *Theory of Functions* by Titchmarsh [11]. Squires further clarifies in a footnote that λ can be taken to be less than $\frac{\pi}{2}$ and can even be taken *negative*.

In order to apply Carlson's theorem to the problem of analytic continuation of the partial-wave amplitudes to complex l -plane, one first defines a function $a(l, s)$ which satisfies, for some L , the following:

(a) $a(l, s) = a_l(s)$, $l = \text{positive integer which is greater or equal to } L$ (which need not be an integer)

(b) $a(l, s) = O(e^{\lambda|l|})$, as $l \rightarrow \infty$ in $\text{Re } l \geq L$, with $\lambda < \pi$

(c) $a(l, s)$ is holomorphic in $\text{Re } l \geq L$.

Condition (b) should be carefully distinguished from the Lehmann large- l asymptotics $a_l(s) \simeq e^{-\Gamma l}$ as $l \rightarrow \infty$; the latter is for real positive integral l only, whereas the former holds for complex l in the specified range. Now one applies Carlson's theorem for $a(l, s) - a_l(s)$ along with the condition (b) for $a(l, s)$ and the Lehmann asymptotics for $a_l(s)$ to establish the desired analytic continuation.

It is important to raise the question as to whether the conditions (a), (b), (c), and consequently the existence of such an $a(l, s)$ can at all be realized. This is addressed in the next subsection.

13.3.1 Cases where Angular Momentum can be Complexified

The function $a(l, s)$ defined as the analytic continuation of the partial-wave amplitude $a_l(s)$ to complex values of l had to satisfy the conditions (a), (b), (c) spelt in the section above in order that Carlson's theorem would guarantee the extension to exist and be unique. Condition (a) is in itself not as restrictive as the conditions (b) and (c).

There are two circumstances where an $a(l, s)$ satisfying these conditions definitely exists. These are (i) Potential scattering in Quantum Mechanics, and, (ii) relativistic S-matrix satisfying the Mandelstam Representation (see Sect. (11.7)). Squires gives a detailed account of the aspects of Potential Scattering where this is manifest in his Chap. 2. In this chapter Squires discusses how conditions (b) and (c) can be fulfilled in Potential Scattering as well as various aspects of analyticity of $a(l, s)$ ($a(l, k)$ in Potential Scattering). A remarkable feature of Potential Scattering is that Mandelstam Representation can be proved there. The use of Potential Scattering to motivate and gain a deeper understanding of the analytic continuation of scattering amplitudes to the complex plane owes it to the pioneering researches of Tulio Regge [12]. The idea of *Regge Poles*, which will play a central role in this book, and which will be elaborated upon in a later section, was also introduced by Regge [1] as well as his collaborators Bottino and Longini [13]. A very nice account of these developments may be found in the recent review by Bottino [14]. Obviously, there were many other pioneering contributions as this became an area of very intense research. The books of Squires [10] and Frampton [9] both have a large number of references to them. Particularly important and impressive among them are the papers by Blankenbecler and Goldberger [15] who investigated in great detail the behaviour of scattering amplitudes at high energies both in potential scattering and in field theory, and, by Virendra Singh [16] whose treatment of Coulomb Scattering in both non-relativistic and relativistic quantum theory very explicitly demonstrates all these ideas. Singh works out the consequences in both Klein-Gordon and Dirac relativistic quantum theories.

In Chap. 3, Squires discusses how in relativistic S-matrix theory, the above-mentioned continuation to $a(l, s)$ exists if Mandelstam Representation could be assumed. As already mentioned, in potential scattering the Mandelstam Representation can actually be proved. Actually, on closer examination of the arguments presented by Squires in his Chap. 3, it appears as if just the existence of fixed- s dispersion relations is enough to prove the existence of $a(l, s)$. While the Mandelstam Representation remains unproven even to this day, we saw that fixed- t dispersion relations could be proved within the ambit of relativistic quantum field theory with reasonable confidence. That can be done for fixed- s dispersion relations too.

The single-variable dispersion relations at fixed- s are of the form:

$$A(s, t) = \frac{1}{\pi} \int_{t_0} dt' \frac{A_t(s, t')}{t' - t} + \frac{1}{\pi} \int_{u_0} du' \frac{A_u(s, u')}{u' - u} \quad (13.12)$$

Here A_t, A_u are respectively the t -channel and u -channel discontinuities. We will now try to clarify whether Mandelstam Representation is really needed for $a(l, s)$ to exist. Recall that Mandelstam Representation would demand that the discontinuities $A_t(s, t), A_u(s, u)$ be further specified as

$$\begin{aligned} A_t(s, t) &= \frac{1}{\pi} \int_{s_0} ds' \frac{\rho_{st}(s' t')}{s' - s} + \frac{1}{\pi} \int_{u_0} du' \frac{\rho_{tu}(t' u')}{u' - u} \\ A_u(s, u) &= \frac{1}{\pi} \int_{s_0} ds' \frac{\rho_{su}(s' u')}{s' - s} + \frac{1}{\pi} \int_{u_0} du' \frac{\rho_{tu}(t' u')}{t' - t} \end{aligned} \quad (13.13)$$

where $\rho_{st}, \rho_{tu}, \rho_{su}$ are the spectral functions characterizing the double spectral representations. The thresholds s_0, t_0, u_0 are determined by unitarity for the respective channels. According to Squires, the spectral functions must be *uniformly bounded* according to which all of them are bounded by the *same power* of one variable as the other variable tends to infinity. This immediately implies that for all fixed s , there exists a λ such that

$$A_t(s, t) = 0(t^\lambda) \quad A_u(s, u) = 0(u^\lambda) \quad (13.14)$$

as t , and hence u , tend to infinity. This also means the same number N , where N is the smallest integer greater than λ , of subtractions should suffice to make both the integrals, i.e. over $A_t(s, t')$ and $A_u(s, u')$ converge.

How would this change if only fixed- s dispersion relations are invoked but not the Mandelstam Representation itself? Then there is no particular reason why the dt', du' integrations in Eq. (13.12) would require the *same* number of subtractions, though that possibility is not ruled out. That may happen due to additional symmetry requirements. If the subtractions are indeed the same, the full Mandelstam Representation would not have any real advantage over just using the fixed- s dispersion relations. We also saw in an earlier chapter that it is sometimes possible to put limits on the number of subtractions itself, as was done by Jin and Martin [17].

If on the other hand, the number of subtractions are indeed *different*, then there is the possibility that Mandelstam Representation and fixed- s dispersion relations will work differently in the existence of $a(l, s)$. We now investigate this generic situation and also explain the significance of the number of subtractions to the complex angular momentum approach. For this, let us start with *unsubtracted* dispersion relations of the type

$$A(\nu, t) = \frac{1}{\pi} \int d\nu' \frac{\text{Im } A(\nu', t)}{\nu' - \nu} \quad (13.15)$$

As explained before, the need for subtractions arises from the integrals not converging because of unfavourable growth of $\text{Im } A(\nu', t)$ at very large ν' . Suppose the value of $A(\nu, t)$ at $\nu = \nu_1$ is the *function* $A(\nu_1, t)$. Then, it readily follows that

$$A(\nu, t) = A(\nu_1, t) + \frac{1}{\pi} \int d\nu' \frac{\text{Im } A(\nu', t)}{\nu' - \nu} \frac{(\nu - \nu_1)}{(\nu' - \nu_1)} \quad (13.16)$$

Now the integral is more likely to converge because of the additional factor $\frac{1}{\nu' - \nu_1}$. This of course comes at the price of supplying the function $A(\nu_1, t)$ as *additional data*. It is easy to generalize these considerations to cases where, say, r subtractions are necessary (the following should only be used in cases where more than one subtraction is necessary, to avoid some ambiguities):

$$A(\nu, t) = \sum_{i=1}^r A(\nu_i, t) \prod_{j \neq i, j=1}^r \frac{(\nu - \nu_j)}{(\nu_i - \nu_j)} + \frac{1}{\pi} \int d\nu' \frac{\text{Im } A(\nu', t)}{\nu' - \nu} \prod_{j=1}^r \frac{(\nu - \nu_j)}{(\nu' - \nu_j)} \quad (13.17)$$

(see Eq. (1.45) of [9]). In Frampton's analysis, $A(\nu, t)$ is the amplitude for t -channel scattering, so t is like the centre of mass energy variable and ν the invariant corresponding to the scattering angle in the t -channel. Explicitly stated, $z_t = \cos \theta_t = \frac{2\nu}{t - 4m^2}$ (this is for equal mass scattering) and $\nu = s - u$. Now we carry out a partial-wave analysis of $A(\nu, t)$ wrt to the scattering angle z_t . We shall treat the subtraction "constants", and the subtracted integrals separately to bring out their significance more clearly. The partial-wave amplitudes are given by

$$a_l(t) = \frac{1}{2} \int_{-1}^{+1} dz_t A(t, z_t) \quad (13.18)$$

where we have traded the variable ν for z_t . Because of the $j \neq i$ condition, the term involving the "subtraction polynomials" $A(\nu_i, t)$ are polynomials in ν , and hence z_t , of maximal degree $r - 1$. Their contribution to the partial-wave expansion will therefore be of the form

$$\sum_{l=0}^{r-1} (2l + 1) a_l(t) P_l(z_t) \quad (13.19)$$

We follow Squires (see Eq. (3.8) of [10]) in treating the integral part of the multiply subtracted amplitude; we first note that each factor $\frac{(\nu - \nu_j)}{(\nu' - \nu_j)}$ can be recast as $\frac{(z_t - z_j)}{(z'_t - z_j)}$ because of the linear relationship between ν and z_t . Now in order to carry out the partial-wave analysis, one substitutes the integral part of Eq. (13.17) in Eq. (13.18), interchanges the order of z_t and ν' (which is permitted). One then has to perform the angular integral

$$\int_{-1}^{+1} dz_t \frac{P_l(z_t)}{z'_t - z_t} \prod_{j=1}^r \frac{z_t - z_j}{z'_t - z_j} = \int_{-1}^{+1} dz_t P_l(z_t) \left\{ \frac{1}{z'_t - z_t} + \sum_{n=0}^{r-1} c'_n z_t^n \right\} \quad (13.20)$$

where the coefficients c'_n do not depend on z_t but can depend on all the z_i as well as on z'_t . Now the very important point is that the polynomial part involving these c'_n does not contribute to the integral for $l \geq r$! Since

$$\int_{-1}^{+1} dz_t \frac{P_l(z_t)}{z'_t - z_t} = 2 Q_l(z'_t) \quad l = +integer \quad (13.21)$$

The partial-wave amplitudes for $l \geq r$ are given by

$$a_l(t) = \frac{1}{\pi} \int_{z_0} \text{Im } A(t, z'_t) Q_l(z'_t) \quad (13.22)$$

The details of how the partial-wave amplitudes are continued to the complex l -plane, $a(l, t)$, can be found in [9,10]. Now we return to our original problem of s -channel scattering amplitudes, and the fixed- s dispersion relations of Eq. (13.12). Everything we have said so far can be individually applied to the A_t, A_u integrals. The s -channel scattering angle $z_s = \cos \theta_s$ can be represented both as $1 + \frac{t}{s-4m^2}$, and, $-(1 + \frac{u}{s-4m^2})$ (for equal mass case). For the dt' integral one uses the first representation, and for the du' the second.

Recognizing that in principle the number of subtractions for the two integrals could be different, let P be the number of subtractions for the first, and R be the number of subtractions in the second. The subtraction “constants” in the first case give rise to a polynomial in t of degree $P - 1$, which can be recast as a polynomial in z_s of the same degree, and hence in the form $\sum_{i=0}^{P-1} c_i(s) P_i(z_s)$. Likewise, the subtraction constants in the second case give rise to a polynomial in u of degree $R - 1$ which too can be recast as a polynomial in z_s of degree $R - 1$, or equivalently as $\sum_{i=0}^{R-1} d_i(s) P_i(z_s)$. Together, they can be written as $\sum_{i=0}^{N-1} a_i(s) P_i(z_s)$, with N as the larger of (P, R) . The integral contributions can likewise be written as for $l \geq N$

$$a_l(s) = \frac{1}{\pi} \int_{z_0}^{\infty} dz' A_t(s, t') Q_l(z') - \frac{1}{\pi} \int_{z_0}^{\infty} A_u(s, u') Q_l(-z') \quad (13.23)$$

(see Eqs. (3–10) of [10]). These are identical in form to those given by Squires based on the Mandelstam Representation. Thus, as stressed earlier the fixed- t (or fixed- s as the case may be) dispersion relations suffice as far as continuation of partial-wave amplitudes to the complex plane is concerned, and it is not necessary to invoke the as yet unproven Mandelstam Representation.

13.4 The Sommerfeld-Watson Transform

Recall that the conditions (a),(b),(c) to be satisfied by $a(l, s)$ involved an L (not necessarily integer) whose significance was not very clear. However, from the previous section on the existence of $a(l, s)$ in relativistic S-matrix theory based on dispersion

relations, we saw that the partial-wave analysis was naturally split into a sum and an integral. The range of l -values was thus split into two intervals characterized by an integer N which was the larger of the number of subtractions for the dt' , du' integrations.

Choosing a contour C in the complex l -plane such that it encloses all integer l -values larger than $N(L)$, it is possible to rewrite the partial-wave expansion as

$$A(s, t) = -\frac{1}{2i} \int_C dl (2l+1) a(l, s) \frac{P_l(-z)}{\sin \pi l} + \sum_{l=0}^{N(L)} (2l+1) a_l(s) P_l(z) \quad (13.24)$$

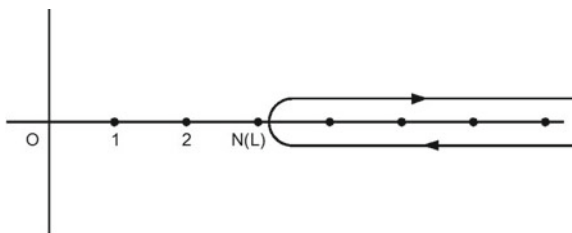
This is called the *Sommerfeld-Watson Transform* and is a technique for converting sums into integrals. But it is different from the Euler-McLaurin formula. It should be carefully noted that in the integral l is a complex variable, while in the sum it is a positive integer. This representation makes critical use of the fact that by condition (c) $a(l, s)$ is holomorphic in the region enclosed by the contour C . The only singularities (poles) arise out of the zeros of $\sin \pi l$ at the positive integer values, and Cauchy residue theorem yields the terms in the partial-wave expansion for $l > N(L)$. One has to invoke the property $(-1)^l P_l(-z) = P_l(z)$ to reproduce the original partial-wave expansion.

Figure 13.1 shows the poles at the integer values of l as well as the contour C : It should be appreciated that the Sommerfeld-Watson transform above is rather delicately put together. In order to have simple poles at integer values of l (which, because of the choice of the contour C are restricted to be positive also), the $\sin^{-1} \pi l$ factor is chosen. Since its residues at $l = n$ are proportional to $(-1)^n$, the Legendre Polynomial with $(-z)$ as argument is chosen. Since $a(l, s)$ is by construction holomorphic in the region of C , and since $P_l(-z)$ in the same region has no singularities in the l -plane, it is only the simple poles at $l = n$ that count.

On making use of the holomorphicity of $a(l, s)$ for $\text{Re } l \geq L$, the contour C can be deformed to a line running along $\text{Re } l = L$ from $-\infty$ to ∞ in the $\text{Im } l$ direction, completed by a semi-circle at infinity in the complex l -plane. The contribution from this semi-circle at infinity can be neglected in view of the asymptotic fall-off of $a(l, s)$ for large $|l|$, as per condition (b), and,

$$\left| \frac{P_l(-z)}{\sin \pi l} \right| < l^{-1/2} \frac{e^{|\text{Im } \theta \cdot \text{Re } l + (\pi - \text{Re } \theta) \cdot \text{Im } l|}}{e^{\pi |\text{Im } l|}} \quad (13.25)$$

Fig. 13.1 The Sommerfeld-Watson transformation



This leads to the representation

$$A(s, t) = -\frac{1}{2i} \int_{L-i\infty}^{L+i\infty} (2l+1)a(l, s) \frac{P_l(-z)}{\sin \pi l} dl + \sum_{l=0}^{N(L)} (2l+1)a_l(s) P_l(z) \quad (13.26)$$

for *physical* $-1 \leq z \leq +1$. It should be strongly emphasized that the above representation is still for physical values of z only.

Now comes the magical bonus from the Sommerfeld-Watson transform as modified to Eq. (13.26)! Since in the integral part $\text{Re } l$ is *fixed* at L , as per Eq. (13.25), the value of $\text{Im } \theta$ is irrelevant for convergence, which is now guaranteed by the additional exponential fall-off due to the factor in the denominator which came at no extra cost, and due to the $\sin \pi l$ factor in the Sommerfeld-Watson transform. Stated differently, the representation of $A(s, t)$ by Eq. (13.26) allows for an analytical continuation into complex θ that is valid for all $\text{Im } \theta$, and not just those restricted to the Lehmann Ellipse. This is the power of the complex angular momentum.

13.5 Regge Poles and Their Properties

Suppose we try to enlarge the domain of holomorphicity of $a(l, s)$ to the *left* of $\text{Re } l = L$ to say $\text{Re } l = L' < L$. If one could succeed in that, the Sommerfeld-Watson representation of $a(l, s)$ would take the same form as before but with L replaced by L' . But there is no guarantee that this can always be done. The next best thing that can happen is that in the region $L' < \text{Re } l < L$, $a(l, s)$ is *meromorphic*, i.e. it is holomorphic apart from *poles* in l . These are called *Regge Poles* named after Tulio Regge who discovered them in his pioneering studies in potential scattering [1]. Again, even within potential scattering there is no guarantee that the new region will only have such poles and not more complicated structures like *Branch Cuts*, etc. In fact, V. Singh in his studies on Coulomb Scattering [16] had found that in the case of Relativistic Coulomb Scattering (both Klein-Gordon and Dirac) that in addition to poles, there were indeed branch cuts. But these were of the kind that did not move with energy. Branch cuts of the latter kind, are called *Regge Cuts*. It is the plethora of such objects that made further investigations in this exciting area very challenging even to hardened experts in complex analysis. We shall restrict attention only to Regge Poles. Though they were discovered in potential scattering, conjecturing their existence in relativistic S-matrix theory has brought an enormous degree of conceptual clarity. As we shall see, Regge Poles play a crucial role in leading to the concept of *Hadronic Strings* first and eventually to *String Theory*.

So Regge Poles are poles in the complex l -plane that move with energy. An s -channel Regge Pole would be a pole in the complex l -plane at $\alpha_i(s)$. As s varies, the position of the pole traces out a trajectory in l -plane called a *Regge Trajectory*. More specifically, the analytically continued $a(l, s)$ will have a contribution of the type

$$a(l, s) = \frac{\beta_i(s)}{l - \alpha_i(s)} + \dots \quad (13.27)$$

$\beta_i(s)$ is called the residue of the Regge Pole at $l = \alpha_i(s)$. The label “ i ” denotes the i th Regge pole. There are two very important aspects of Regge poles both of which have received considerable theoretical and experimental support in relativistic S-matrix theory also, and which will also be seen to lead to the important notion of *Duality*. These are the relation to stable bound states and resonances on the one hand, and, asymptotic behaviour of scattering amplitudes.

Before discussing these, it is important to state the implications of the analytic continuation of $a_l(s)$ to $a(l, s)$, more specifically that of Carlson’s theorem to *elastic unitarity*, i.e. the unitarity relation for the range $4m^2 \leq s \leq s_I$, where s_I is the so-called *inelastic threshold*, that is, the value of s where multiparticle processes set in (for simplicity sake, we continue to discuss scattering of two equal mass scalars). In this range, the partial-wave amplitudes satisfy

$$a_l(s) - a_l^*(s) = \left(\frac{k^2}{k^2 + m^2}\right)^{1/2} 2i a_l(s) a_l^*(s) \quad (13.28)$$

where k is the centre of mass momentum. It then follows immediately from Carlson’s theorem that the continued $a(l, s)$ will satisfy

$$a(l, s) - a^*(l^*, s) = \left(\frac{k^2}{k^2 + m^2}\right)^{1/2} 2i a(l, s) a^*(l^*, s) \quad 4m^2 \leq s \leq s_I \quad (13.29)$$

13.5.1 Bound States and Resonances

Stable Bound States: The implication of this continued version of unitarity is that $a(l, s)$ for $s < 4m^2$ is real and starts to become imaginary for $s > 4m^2$. Let us first consider the region $s < 4m^2$. Unitarity implies that in this region $\alpha_i(s)$ is real. If for some value s_b , $\alpha_i(s_b)$ is a positive integer, say, l_0 , then this represents a *Stable Bound State*.

Resonances: Resonances are enhancements in physical scattering processes. Hence they correspond to $s > 4m^2$. The width of the resonance is a measure of its lifetime. Hence resonances are *unstable*. It proves useful to introduce the notion of a *Physical Sheet* in the complex s -plane. Unitarity, say in the s -channel, mandates a branch cut in complex s -plane starting at $s = 4m^2$. For values of $s \geq 4m^2$, the physical sheet is taken to be just above this branch cut. Hence $\text{Im } s > 0$. When $\text{Im } s < 0$ it is said to lie on the *Unphysical Sheet*.

Now consider a Regge trajectory such that for $s = s_0 > 4m^2$, $\text{Re } \alpha(s_0) = l_0$ (l_0 is a positive integer). According to unitarity, $\alpha(s)$ will now be complex, but let us consider the situation where $\text{Im } \alpha(s_0)$ is still *small*. The analytically continued $a(l_0, s)$ with the Regge pole $\alpha(s)$

$$a(l_0, s) = \frac{\beta(s)}{l_0 - \alpha(s)} \quad (13.30)$$

will have a pole at

$$s \simeq s_0 + \frac{l_0 - \alpha(s_0)}{\frac{d\alpha(s)}{ds}|_{s_0}} \quad (13.31)$$

(this can be seen by just Taylor expanding $\alpha(s)$ near $s = s_0$). Because $\text{Re } \alpha(s_0) = l_0$, the pole is shifted from s_0 by a small imaginary part. It is equivalent to having complex energy. From temporal development e^{-iEt} only a negative imaginary part to E represents a decaying solution. Therefore, the imaginary part of s above also has to be negative, and the pole has to lie on the unphysical sheet (confusing terminology!). This has to be reached by going downward through the cut. In the language of complex analysis, it is on the *Second Riemann Sheet*. This would be the description of a resonance in the l_0 -th partial wave.

Thus we see that Regge poles unite stable bound states and resonances in a simple, yet very striking manner. Furthermore, as s increases along a Regge trajectory, $\text{Re } \alpha(s)$ can hit other positive integer values. Thus the Regge pole picture naturally groups resonances with the same quantum numbers but different spins.

A remarkable empirical validation of this was provided by Chew and Frautschi [18], who plotted angular momenta of known resonances then against mass (or square of masses for bosons). They did find that resonances lay on different trajectories. Though highly suggestive, it was not possible to make serious claims as most trajectories had two points on them! Nevertheless, some features stood out; one of them being that trajectories to a good degree were *linear*:

$$\alpha(s) = \alpha(0) + \alpha'(0)s \quad (13.32)$$

and, *parallel*! The common slope was found to be close to $1(\text{GeV})^{-2}$. There is a much more recent study of Regge trajectories by Alfred Tang and John Norbury [19]. They used the latest data as available in 2000. They found several trajectories with even four points. They claimed that linearity was violated, but an eye-ball estimate would still support linearity to a good degree. It is worth emphasizing that strictly linear trajectories are incompatible with unitarity, and they must bend over.

13.5.2 Regge Asymptotics

In the above, we had restricted attention to only positive values of s . Now we shall see some equally dramatic consequences of Regge poles, but applicable to $s < 0$. Obviously, this is an unphysical region as far as s-channel scattering is concerned, but can indeed be part of the physical region for t-channel and u-channel processes.

To proceed one moves the contour in the Sommerfeld-Watson transform to the left of the original L (determined by subtractions) to $L' < L$. But L' is still $\geq -\frac{1}{2}$. The reason for this is somewhat technical, and the interested reader is referred to [10]. Let us further assume that in moving to the new region we encounter *two* Regge poles $\alpha_1(s)$, $\alpha_2(s)$ with residues $\beta_1(s)$, $\beta_2(s)$:

$$a(l, s) = \frac{\beta_1(s)}{l - \alpha_1(s)} + \frac{\beta_2(s)}{l - \alpha_2(s)} + \dots \quad (13.33)$$

These poles contribute as we move from L to L' and their contributions can be calculated using Cauchy Residue Theorem, leading to

$$\begin{aligned} A(s, t) = & -\frac{1}{2i} \int_{L'-i\infty}^{L'+i\infty} dl (2l+1) a(l, s) \frac{P_l(-z)}{\sin \pi l} \\ & + \sum_{l=0}^{N(L')} (2l+1) a_l(s) P_l(z) \\ & - \pi \sum_{i=1}^2 (2\alpha_i + 1) \beta_i(s) \frac{P_{\alpha_i}(-z)}{\sin \pi \alpha_i} \end{aligned} \quad (13.34)$$

with $\text{Re } \alpha_i > L'$, $i = 1, 2$. Now we want to investigate the asymptotic behaviour of $A(s, t)$ for very large t , equivalently, large z . One uses the asymptotic behaviour of the Legendre Polynomials for large z [5, p.126]:

$$P_\alpha(z) \rightarrow \frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(\alpha + 1)} \frac{(2z)^\alpha}{\sqrt{\pi}} \quad z \rightarrow \infty \quad (13.35)$$

The asymptotic behaviour of the first two terms of Eq. (13.34) is $t^{L'}$, but those of the last (Regge pole terms) is t^{α_i} . Thus with our restriction $\text{Re } \alpha_i > L'$, the Regge pole terms dominate over the first two terms. If L' had been moved to $< -\frac{1}{2}$, the pole terms would not have dominated! Let us explicitly exhibit the asymptotic behaviour of these:

$$\begin{aligned} A(s, t) \rightarrow & -\sqrt{\pi} \left\{ \frac{2\alpha_1 + 1}{\sin \pi \alpha_1} \frac{\Gamma(\alpha_1 + \frac{1}{2})}{\Gamma(\alpha_1 + 1)} \right\} \beta_1(s) (-2z)^{\alpha_1} \\ & - \sqrt{\pi} \left\{ \frac{2\alpha_2 + 1}{\sin \pi \alpha_2} \frac{\Gamma(\alpha_2 + \frac{1}{2})}{\Gamma(\alpha_2 + 1)} \right\} \beta_2(s) (-2z)^{\alpha_2} \end{aligned} \quad (13.36)$$

If we order the two trajectories so that $\text{Re } \alpha_1 > \text{Re } \alpha_2$, then the α_1 trajectory dominates over the α_2 in the asymptotic limit.

In summary, the high energy behaviour in the t -channel is governed by the Regge poles in the s -channel and if there are many Regge poles in the s -channel, the pole lying to the farthest on the right will dominate. The s -channel Regge pole at $\alpha(s)$ will result in the t -channel behaviour of $t^{\alpha(s)}$ for very large t . There is a caveat, however! If t -channel is the physical channel, $t > 4m^2$ and $s < 0$! But the Sommerfeld-Watson transform was derived for positive s . This means the proof of the Sommerfeld-Watson transform given above needs to be refined. We shall not go into that but refer the interested reader to Sect. 1.5 of [10] where it is shown that as long as single-variable dispersion relations hold, the asymptotic behaviour for large t is as given by the Regge pole form, for all fixed s .

In the coming chapters we shall explore the deep consequences arising out of the interplay between these two aspects of Regge poles in relativistic S-matrix theory. Many very interesting phenomenological consequences of Regge poles are discussed in [9].

References

1. T. Regge, *Nuovo Cim.* **14**, 951 (1959)
2. P. Goddard, in *Memorial Volume for Stanley Mandelstam*, ed. by N. Burkovits (World Scientific, 2017)
3. R.J. Eden, P.V. Landshoff, D.I. Olive, J.C. Polkinghorne, *The Analytic S-matrix* (Cambridge University Press, 1966)
4. H. Lehmann, *Nuovo Cim.* **10**(4), 579 (1958)
5. Batemann Manuscript Project, *Higher Transcendental Function*, vol. 1
6. R. Jost, H. Lehmann, *Nuovo Cim.* **5**, 1598 (1957)
7. F.J. Dyson, *Phys. Rev.* **110**, 1460 (1958)
8. *Summer School in Elementary Particle Physics* (Brookhaven National Laboratory, July 22–Aug 29, 1969)
9. P.H. Frampton, *Dual Resonance Models* (W.A. Benjamin, 1974)
10. E.J. Squires, *Complex Angular Momentum and Particle Physics* (Benjamin, W.A., 1963)
11. E.C. Titchmarsh, *Theory of Functions* (Oxford University Press, 1939)
12. T. Regge, *Nuovo Cim.* **8**, 671 (1958)
13. A. Bottino, A.M. Longoni, T. Regge, *Nuovo Cim.* **23**, 954 (1962)
14. A. Bottino, A retrospective look at Regge Poles (2018), [arXiv:1807.02456](https://arxiv.org/abs/1807.02456) [hep-ph]
15. R. Blankenbecler, M.L. Goldberger, *Phys. Rev.* **126**, 766 (1962)
16. V. Singh, *Phys. Rev.* **127**, 632 (1962)
17. Y.S. Jin, A. Martin, *Phys. Rev. B.* **135**, 1375 (1964)
18. G.F. Chew, S.C. Frautschi, *Phys. Rev. Lett.* **7**, 394 (1961); *Phys. Rev. Lett.* **8**, 41 (1962)
19. A. Tang, J.W. Norbury, Properties of Regge Trajectories (2000), [arXiv:0004078v3](https://arxiv.org/abs/0004078v3) [hep-ph]

Superconvergence Relations, FESR and Duality

14

14.1 Introduction

This chapter explores one of the deepest concepts to emerge from analyticity considerations, namely, that of *Duality*. We shall go into enough details to enable readers to appreciate the simple but beautiful arguments that led to Duality. Once again, it is practically impossible to discuss the many exciting developments and works of individuals that culminated in what may easily be called a breakthrough in the analytic S-matrix programme. One may recall that the early applications of Dispersion Relations were limited in their success as the imaginary part of the scattering amplitude was to be used as an input via the *Optical Theorem* relating it to the observable data on the total cross sections. It was essentially *data-driven*, and it was severely constrained by the reliability of available data. On the theoretical side too, it did not yield any insights into the nature of strong interactions. What changed this substantially are the new insights provided by Regge Trajectories.

14.2 Superconvergence Relations

Consider an amplitude $A(\nu, t)$ analytic in ν , where t is the Mandelstam invariant related to the scattering angle in the direct channel (say, s -channel), and ν is either $t - u$, or, the laboratory energy as the case may be. Let us consider cases where $A(\nu, t)$ falls off faster than ν^{-1} for asymptotically large ν . Clearly, not all amplitudes will satisfy this restriction, and we shall discuss what to do in those more general cases, later.

A fall-off faster than ν^{-1} ensures that *both* $A(\nu, t)$ and $\nu A(\nu, t)$ satisfy *unsubtracted* dispersion relations. Consequently (see also, Frampton's book [1] Sect. 1.5),

$$\begin{aligned} A(\nu, t) &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{d\nu'}{\nu' - \nu} \text{Im } A(\nu', t) \\ \nu A(\nu, t) &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{d\nu'}{\nu' - \nu} \nu' \text{Im } A(\nu', t) \end{aligned} \quad (14.1)$$

Multiplying the first by ν and subtracting from the second, one gets the *Superconvergence Relation*:

$$\int_{-\infty}^{\infty} d\nu' \text{Im } A(\nu', t) = 0 \quad (14.2)$$

While it is trivially satisfied when $A(\nu, t)$ is an *odd* function of ν , for *even* functions it leads to

$$\int_0^{\infty} d\nu' \text{Im } A(\nu', t) = 0 \quad (14.3)$$

Though the example above is applicable to scattering of *Spin-less* particles, these considerations can be applied to arbitrary scattering processes as long as the requisite fall-off asymptotically is assured. For example, in the case of Pion-Nucleon scattering, due to spin and isospin considerations there are actually *four* independent amplitudes $A^{(\pm)}(s, t)$, $B^{(\pm)}(s, t)$ [1,2]. Superconvergence relations are in principle applicable to all of them as long as their asymptotic fall offs are appropriate.

Historically, such superconvergence relations were first obtained by Alfaro, Fubini, Rossetti and Furlan [3]. Their derivations were based on *Current Algebra* techniques, and the superconvergence of the amplitudes arose due to some algebraic properties of the amplitudes. The authors then went on to point out that such superconvergence relations (called “strong interaction sum rules” by them) could be due to rather general grounds like analyticity, unitarity and high energy behaviour.

At this stage the superconvergence relations only hint at some broad connections between the amplitudes at low and high energies and are not more specific than that. However, Alfaro et al. had already pointed out some possible usefulness of such sum rules. For example, in the case of $\rho\pi$ scattering, assuming that the amplitudes are saturated by the π , ω , ρ poles, they showed that the sum rules could lead to relations like (various coupling constants had differing mass dimensions)

$$(g_{\omega\rho\pi}^2 + g_{\phi\rho\pi}^2)m_\rho^2 - 4g_{\rho\pi\pi}^2 = 0 \quad (14.4)$$

Frampton [1] also points out that a narrow resonance approximation to the amplitudes leads to such relations between hadron masses and coupling constants.

14.2.1 Igi's Significant Next Step

The derivation of the superconvergence relations hinged on a rather restrictive high energy behaviour, i.e. the amplitude must fall-off *faster* than ν^{-1} . Quite obviously, amplitudes do not obey this in general. Regge asymptotics offers some handle on this question. For particles with spin, the Regge asymptotic behaviour gets modified to [1]

$$A_h(s, t) \rightarrow s^{\alpha(t)-h} \quad (14.5)$$

where $\alpha(t)$ is the t -channel trajectory, and h is the t -channel helicity flip. Therefore, with spin and sufficiently high h , the conditions for superconvergence relations can indeed be met. Even for particles without spin, if the asymptotic behaviour is dominated by Regge behaviour (a priori, there is no reason this has to be so), and if the leading trajectory has $\alpha < -1$, then too the superconvergence relations must hold. But even these are not generally true. What if the asymptotic behaviour is dominated by Regge asymptotics, but with $\alpha > -1$? In that case, the conditions for the validity of superconvergence relations obviously do not hold.

Igi [4] was the first to show a way out. Though conceptually simple and easy to implement, this was a significant step forward which opened the ways for very important developments like *Finite Energy Sum Rules* (FESR), and Duality. His basic premises were that all Baryons and Mesons are associated with Regge poles and that they control asymptotic behaviour in crossed channels (see Chap. 13 for a detailed discussion of Regge Poles and their properties).

He investigates Pion-Nucleon *non-charge exchange* processes. Because of the non-charge exchange nature, in the t -channel the so called *Pomeron* trajectory will dominate. This is described by $\alpha_P = 1$. Igi also restricts his analysis to the forward direction. The relevant amplitude is then

$$f^{(+)}(\nu) = \frac{1}{4\pi} [A^{(+)}(\nu) + \nu B^{(+)}(\nu)] \quad (14.6)$$

He chooses ν to be the laboratory energy which is also the cosine of the scattering angle in t -channel at $t = 0$. The main idea of Igi is to separate the Pomeron contribution from the amplitude and write a superconvergence relation for the rest. At first, he assumes that there are no other trajectories between the Pomeron and $\alpha = -1$ in which case superconvergence should indeed hold for the difference. In case this is not so, his method can easily be extended by subtracting the relevant additional trajectories also, leading to definite corrections. In fact he proposes that as the way to experimentally check for the presence of such additional singularities.

With the Pomeron being the only source of a fall-off faster than ν^{-1} , Igi splits the total amplitude as

$$f^{(+)}(\nu) = F_P(\nu) + f^{(+)'}(\nu) \quad F_P(\nu) = \frac{\beta_P}{\pi} \nu \ln \frac{1-\nu}{1+\nu} - \frac{2\beta_P}{\pi} \quad (14.7)$$

Under the assumptions made, it is clear that $f^{(+)\prime}(\nu)$ vanishes fast enough for very large ν for it to satisfy an *unsubtracted* dispersion relation:

$$f^{(+)\prime}(\nu) = \frac{1}{4\pi} \frac{g_r^2}{2M} \nu_B \left[\frac{1}{\nu_B - \nu} + \frac{1}{\nu_B + \nu} \right] + \frac{1}{\pi} \int_1^\infty d\nu' \text{Im} f^{(+)\prime}(\nu') \left[\frac{1}{\nu' - \nu} + \frac{1}{\nu' + \nu} \right] \quad (14.8)$$

Here $\nu_B = -\frac{\mu^2}{2M}$ with μ, M being the pion and nucleon masses, respectively. The nucleon pole is separated for convenience. The imaginary part is related to the total cross section $\sigma_{tot}(\nu) = \frac{1}{2} \cdot (\sigma_{tot}^{\pi^+ p}(\nu) + \sigma_{tot}^{\pi^- p}(\nu))$ by the Optical theorem:

$$\text{Im} f^{(+)\prime}(\nu) = \frac{\sqrt{\nu^2 - 1}}{4\pi} \sigma_{tot}(\nu) - \beta_P \nu \quad (14.9)$$

From this follows

$$\beta_P = \frac{\sigma_{tot}^{(+)}(\infty)}{4\pi} \quad (14.10)$$

If these total cross sections are not constant at very high energies, this treatment would need refinement. Taking the real part of the above, Igi derives

$$\text{Re} f^{(+)\prime}(\nu) = -\frac{f^2}{M} \frac{1}{\nu^2 - \nu_B^2} + \frac{1}{2\pi^2} P \int_1^\infty d\nu' \left\{ \frac{\nu' \sqrt{\nu'^2 - 1}}{\nu'^2 - \nu^2} \sigma_{tot}^{(+)\prime}(\nu') - \frac{\nu'}{\sqrt{\nu'^2 - 1}} \sigma_{tot}^{(+)}(\infty) \right\} \quad (14.11)$$

where $f^2 = \frac{g_r^2}{4\pi}$. Evaluating this at $\nu = \mu$ and changing integration variable to k' with $k'^2 = \nu'^2 - 1$, Igi arrives at a sum rule for the s-wave scattering length $a^{(+)}$:

$$\left(1 + \frac{1}{M}\right) a^{(+)} = -\frac{f^2}{M} \frac{1}{1 - \frac{1}{4M^2}} + \frac{1}{2\pi^2} \int_0^\infty dk' [\sigma_{tot}^{(+)}(k') - \sigma_{tot}^{(+)}(\infty)] \quad (14.12)$$

It should be noted that all quantities of mass dimension 1 have been scaled wrt pion mass.

Igi goes on to discussing how this scattering length sum rule is to be vindicated experimentally. We skip those details and refer the reader to the original paper. In a strict sense, this sum rule is not exactly of the superconvergence relation we started with, but is very close in spirit to it. Igi could have also derived that without much effort. What is of main value here is the novel way of handling asymptotic rise through subtractions.

14.3 Finite Energy Sum Rules(FESR)

14.3.1 Horn-Schmid Formulation

The next important development was ushered in by the works of Horn and Schmid [5], and of Dolen, Horn and Schmid [6]. The latter was an elaboration of the basic ideas of the former. Both the works are remarkable in terms of the clarity of exposition. Let us start with the former, entitled *Finite Energy Sum Rules*. They motivate their approach by observing that the superconvergence relations on the one hand are limited to amplitudes with sufficiently rapid fall-off asymptotically, and on the other hand are not so practical to implement as the ν -integration extends to infinity. Even if one relies only on data to validate them, data is only available at finite energies, and in those early years, not at very high energies either.

Horn and Schmid propose to tackle these issues as follows: they follow Igi's prescription to subtract the contributions of all Regge trajectories whose asymptotic fall-off is faster than dictated by the superconvergence relations, and in addition, cut-off the integrations at some suitably chosen values. They implicitly assume that the asymptotic behaviour of all amplitudes is governed only by Regge trajectories, and that amplitudes satisfy dispersion relations (analyticity). They discuss how the superconvergence relations are recovered when the energy cut-off is removed in those cases where the asymptotics are dominated by trajectories with $\alpha < -1$.

To bring out the nuances of their ideas more clearly let us look at their two steps separately. For those cases for which *only* the leading Regge trajectory satisfies $\alpha > -1$ but all others are $\alpha_i < -1$, they follow Igi's prescription and conclude $F(\nu) - F_{\alpha > -1}(\nu)$ satisfies superconvergence relations. Horn and Schmid make one more choice of convenience: they replace the characteristic $Q_{-\alpha-1}(z)$ for Regge poles (see Chap. 13) by the far simpler ν^α ; the two forms have the same high energy asymptotics but differ from each other at finite energies. Many workers have preferred this replacement. Thus the Regge trajectory contribution in Horn and Schmid takes the form

$$R(\nu) = \frac{\beta (1 - e^{-i\pi\alpha})}{\sin \pi\alpha \Gamma(\alpha + 1)} \nu^\alpha \quad (14.13)$$

Interestingly, $R(\nu)$ satisfies an *unsubtracted* dispersion relation on its own for $-1 < \alpha < 1$:

$$R(\nu) = \frac{2\nu}{\pi} \int_0^\infty d\nu' \frac{\beta}{\Gamma(\alpha + 1)} \frac{\nu'^\alpha}{\nu'^2 - \nu^2} \quad (14.14)$$

Consequently, $F(\nu) - R(\nu)$ also satisfies a superconvergence relation, i.e.

$$\int_0^\infty d\nu (F(\nu) - R(\nu)) = 0 \quad (14.15)$$

Next they expand on the classes of Regge trajectories that may enter the discussion. They divide them into three classes: (i) all trajectories with $\alpha_i > -1$; these should be subtracted à la Igi, (ii) poles exactly at -1, i.e. $\alpha_k = -1$, these just contribute

a residue β_k to the r.h.s, and finally, (iii) all poles below -1 , i.e. $\alpha_j < -1$; these will not contribute to the superconvergence relations but they do contribute to the FESRs. Let us first write down the full superconvergence relations (for simplicity, let us assume only one trajectory of type (ii)):

$$\int_0^\infty d\nu \{ \text{Im } F(\nu) - \sum_{\alpha_i > -1} \frac{\beta_i}{\Gamma(\alpha_i + 1)} \nu^{\alpha_i} \} = \beta_k \quad (14.16)$$

Now it is to be noted that each term on the l.h.s individually *diverges*! Horn and Schmid sought a formulation where each term is manifestly *convergent*. This they achieved by cutting off the ν -integration at some $\nu_{\max} = N$. N itself was left arbitrary. But now, even the category (iii) trajectories will contribute, as their contribution vanished only when the ν -integration is carried out all the way to infinity. Indeed, what one gets is

$$\int_0^N d\nu \{ \text{Im } F - \sum_{\alpha_i > -1} \frac{\beta_i}{\Gamma(\alpha_i + 1)} \nu^{\alpha_i} \} + \int_N^\infty d\nu \sum_{\alpha_j < -1} \frac{\beta_j}{\Gamma(\alpha_j + 1)} \nu^{\alpha_j} = \beta_k \quad (14.17)$$

This is a *Finite Energy Sum Rule*(FESR). Every term in this equation is obviously finite, and one has better control over all of them. Another noteworthy feature is that all Regge trajectories contribute to it.

Because of the finiteness of each term, they were able to analyse how exactly the usual superconvergence relations would be recovered in the limit N becomes very large. Horn and Schmid go on to consider applications of such FESR. They look at πN -charge exchange reaction. They show how making use of low energy data, one can determine the parameters of the Regge terms dominating high energy behaviour. In the particular example of pion-nucleon charge exchange reaction, they extract the parameters of the rho-trajectory. They also touch upon the issues of the contributions from other Regge trajectories.

The idea of using low energy data to estimate the parameters of high energy Regge behaviour, or more generally the interlinking of low and high energy behaviours makes the beginnings of the idea of *Duality*, and this became the central theme of a large body of research in the coming years.

14.3.2 Dolen, Horn, Schmid Elaboration

These ideas were considerably elaborated by Dolen, Horn and Schmid in a subsequent work [6]. Though the main ideas were the same as in [5], there were many additional details and applications here. An explanation is also given here for why it is preferable to use a power law representation ν^α for the Regge poles instead of $Q_{-\alpha-1}$; the power law expansion of the latter would necessitate additional Regge poles for consistency.

Here too the theme of using low energy data to determine Regge parameters is discussed at great length. Even the use of high energy data to the determination

of low energy parameters is considered, as also the possibility of better Regge fits with both low and high energy data. In short, the interconnectedness of low and high energy behaviours is put on a much stronger foundation. More concretely, they make a prediction of the mass of the ρ -meson.

A most significant feature of this paper is the elucidation of much of the earlier confusions regarding the so-called *Interference Model*. Two distinctive features of scattering amplitudes are (i) low energy region dominated by resonances, and possibly a non-resonant background, and, (ii) high energy region dominated by Regge poles. An early attempt at synthesizing the two apparently disparate features was the so-called *Interference Models* according to which

$$F_{int} = F_{res} + F_{Regge} \quad (14.18)$$

It was also usual to consider only one Regge trajectory. Dolen, Horn and Schmid make the very important observation that interference models of this type actually contradict FESR. To this end, they consider *non-flip* elastic amplitudes at $t = 0$ where $\text{Im } F > 0$. They then choose N large enough that the ν^α behaviour is a good approximation. Then, according to the Interference model, the smoothed out amplitude below N must be larger than ν^α as the resonances contribute only positively. On the other hand, according to FESR, $\text{Im } F$ in this region must average out to ν^α , thus leading to a contradiction.

So these authors argued that the Interference Model suffers from *Double Counting*. In fact, they argued that the smoothed out resonance behaviour is already contained in the Regge terms. They argued that the correct picture is

$$F = F_{res} + F_{Regge} - \langle F_{res} \rangle \quad (14.19)$$

leading to their curious observation that when the resonances are strongly overlapping so that $F_{res} = \langle F_{res} \rangle$,

$$F \simeq F_{Regge} \quad (14.20)$$

even when F_{res} is large by itself! These very perceptive observations about the failure of the interference models were largely responsible for laying a *quantitative* foundation for Duality. It is worth emphasizing that in all these treatments resonances were treated as being very narrow.

Another important aspect highlighted by them is that FESR really play the role of *consistency* relations on Regge Analysis. To appreciate this, one may think of viewing the resonance contributions as due to Regge poles in the direct channel. Then FESR relate Regge poles in direct and crossed channels. This is complementary to using FESR to relate low energy data and Regge parameters. This is a new kind of *Bootstrap* (consistency conditions), which may be called *FESR Bootstrap*. Many of the conclusions of [6] were also reached independently by Logunov, Soloviev and Tavkhelidze [7]. The Bootstrap idea was also independently proposed and thoroughly investigated by [8–10].

14.3.3 Igi-Matsuda FESR

Igi and Matsuda [11] further investigated Superconvergence Relations along similar lines, with the objective of shedding further light on the complex angular momentum plane. They looked at $\pi^- p \rightarrow \pi^0 n$ charge exchange reaction. The so-called *diffraction shrinkage* at high energies had been explained on the basis of a single ρ -trajectory. However, the observed *polarization* characteristics could not be explained by ρ -exchange. Among other proposals, additional trajectories like ρ' had been suggested.

Therefore Igi and Matsuda used FESR techniques to probe the existence of such additional Regge trajectories. They considered the *non-flip* amplitudes in the forward direction. This is described by the combination (see [2] for a good description of the πN -scattering):

$$f^{(-)}(\nu) = \frac{1}{4\pi} [A^{(-)}(\nu) + \nu B^{(-)}(\nu)] \quad (14.21)$$

(being in the forward direction, the Mandelstam variable t is fixed at 0. Following the authors, let us first consider the case where ρ -trajectory is the only one exchanged. As in our earlier discussions, Igi and Matsuda also separate out the ρ -contribution and write (μ is the pion mass)

$$f^{(-)'}(\nu) = f^{(-)}(\nu) - f_\rho(\nu) \quad f_\rho(\nu) = -\beta_\rho \frac{P_{\alpha_\rho}(-\frac{\nu}{\mu}) - P_{\alpha_\rho}(\frac{\nu}{\mu})}{2 \sin \pi \alpha_\rho} \quad (14.22)$$

Note that these authors use the full form of a Regge pole contribution and not use only the power law truncation. If ρ -trajectory is indeed the only singularity in the complex J -plane, $f^{(-)'}(\nu)$ will satisfy an unsubtracted dispersion relation, as does $\nu f^{(-)'}(\nu)$. Separating out the nucleon pole, one writes down the dispersion relation (with $\nu_B = \frac{\mu^2}{2M}$):

$$\begin{aligned} f^{(-)'}(\nu) &= \frac{g_r^2}{4\pi} \frac{\nu_B}{2M} \left(\frac{1}{\nu_B - \nu} - \frac{1}{\nu_B + \nu} \right) \\ &+ \frac{1}{\pi} \int_\mu^\infty d\nu' \left(\frac{1}{\nu' - \nu} - \frac{1}{\nu' + \nu} \right) \text{Im } f^{(-)'}(\nu') \end{aligned} \quad (14.23)$$

Once again, $\text{Im } f^{(-)'}(\nu)$ is given by the Optical Theorem:

$$\text{Im } f^{(-)'}(\nu) = \frac{\sqrt{\nu^2 - \mu^2}}{4\pi} \frac{1}{2} [\sigma_{\pi^- p}(\nu) - \sigma_{\pi^+ p}(\nu)] - \frac{\beta_\rho}{2} P_{\alpha_\rho}(\nu/\mu) \quad (14.24)$$

In arriving at this the authors made use of the *crossing symmetry*, $\text{Im } f^{(-)'}(\nu) = \text{Im } f^{(-)'}(-\nu)$. The resulting superconvergence relation reads:

$$-\frac{g_r^2}{4\pi} \left(\frac{\mu}{2M} \right)^2 + \frac{1}{\pi} \int_\mu^\infty d\nu' \text{Im } f^{(-)'}(\nu') = 0 \quad (14.25)$$

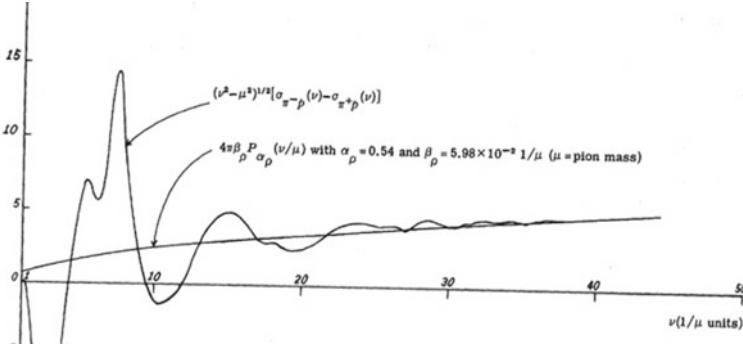


Fig. 14.1 Plot highlighting the resonance and Regge Contributions. Reprinted from [11] ©1967 American Physical Society. Reproduced with permissions. All rights reserved

which the authors recast as (with $f^2 = \frac{g_\rho^2}{4\pi} (\frac{\mu}{2M})^2$)

$$4\pi f^2 - \frac{1}{2\pi} \int_{\mu}^{\infty} d\nu' \{ \sqrt{\nu'^2 - \mu^2} (\sigma_{\pi^- p}(\nu') - \sigma_{\pi^+ p}(\nu')) - 4\pi \beta_{\rho} P_{\alpha_{\rho}}(\nu'/\mu) \} = 0 \quad (14.26)$$

The authors wish to probe the J-plane singularities with relationships of this kind which they call *Sum Rules*. Once again, it is worth emphasizing that each term on its own diverges. It was precisely for keeping control of each term on its own that FESRs were introduced. The authors give details of how they use existing data along with some choices for α_{ρ} , β_{ρ} in some cases, or by fitting data to obtain β_{ρ} for different choices of α_{ρ} . They find the sum rule to be well satisfied and conclude that there is no evidence for other J-plane singularities. They state that their sum rule can easily be extended to include additional J-plane singularities. They claim that the values of $\alpha_{\rho'}$, $\beta_{\rho'}$ proposed in the literature to account for polarization also do not fit the sum rule. A plot of the integrands above has been given by Igi and Matsuda, which we have shown in Fig. 14.1. While the very high energy behaviour of $\text{Im } f$ is indeed well captured by the ρ -trajectory, what is most remarkable is that even the low energy behaviour “on the average” is well approximated by the Regge pole! This is of course only on the average as without averaging there are wild fluctuations coming from the resonances. Roughly speaking, this is strongly hinting at a deep connection between low energy and high energy behaviour, what eventually got to be called Duality. There is no mention of that phrase in the Igi-Matsuda paper itself.

14.4 Alarm Bells Regarding FESR

In all the discussions so far, either only one Regge trajectory, or a few have been used. Attempts have then been made to relate the low energy data and the Regge parameters. When one tries to account for all the low energy resonances as arising out of a single or a few direct-channel resonances, one would have FESR-bootstrap.

In this section, we shall discuss two very important works that point to theoretical limitations to the FESR programme, and the consequent difficulties posed for any reliable estimation of Regge parameters from low energy data. The two works we have in mind are of (i) Mandula and Slansky [12] where they show that when due attention is paid to the t -dependence of trajectory functions, it is impossible to satisfy FESR with a finite number of trajectories, and, (ii) the work of Fujisaki [13] that even with a single leading Regge trajectory, an infinite number of so-called *Daughter Trajectories* will be needed to self-consistently satisfy the FESR. We take up these works one by one now.

14.4.1 The Mandula-Slansky Work

Following Horn and Schmid [5], and, Dolen, Horn and Schmid [6], Mandula and Slansky consider FESR in the sense of relating finite integrals of $\text{Im } A$ to crossed-channel Regge poles. They pay due attention to the t -dependence of Regge parameters so that these FESRs take the form:

$$\frac{1}{\pi} \int_0^s ds' \text{Im } A(t, s') \simeq \sum_i \beta_i(t) s^{\alpha_i(t)+1} \quad (14.27)$$

Since this is a FESR the sum over Regge trajectories includes $\alpha_i < -1$, $\alpha_i = -1$, $\alpha_i > -1$ as explained clearly in [5,6]. These have to hold for *all* values of (s, t) . Thus, there are really an infinite number of FESR relations, and their main point is that with a finite number of Regge trajectories, it is impossible to satisfy all of them, as long as there exists a non-zero minimum gap between the trajectories. More specifically, they show that rising-trajectory models of direct-channel singularities cannot give Regge Asymptotic behaviour. Therefore, they conclude that many applications of FESR in the literature are *inconsistent*. Titling their paper *Misuses of Finite Energy Sum Rules*, they go to the extent of stating their results as a *A Sad Theorem*:

Finitely spaced Regge trajectories of resonances do not yield Regge asymptotic behaviour.

They first consider a simple model with a single trajectory to illustrate their reasoning, and they then extend the same to two-body scatterings with any number of finitely spaced trajectories. The former is the elastic scattering of two identical neutral spin-less particles. They assume that in any channel the amplitude is given by the sum of an infinite number of narrow resonances, all of which lie on a *single* Regge trajectory. For this simple model they further assumed that Regge asymptotic behaviour in each channel to be due to the *same* trajectory in the crossed channel.

Then the *discontinuity* of $A(s, t)$ across s -channel is given by

$$[A(s, t)]_s = \sum_n (2\alpha(s_n) + 1) \bar{\beta}(s_n) \delta(s - s_n) P_{\alpha(s_n)} \left(1 + \frac{2t}{s - 4m^2}\right) \quad (14.28)$$

It should be carefully noted that the residues $\bar{\beta}(s_n)$ have been distinguished from the residue function β that occurs on the r.h.s. of Eq. (14.27) (in this example, i takes only one value). The important point made by the authors is that while the r.h.s is a smooth function of s , the l.h.s is a sum of delta functions, and the relation cannot hold exactly, even in the $s \rightarrow \infty$ limit. They then go on to state that Regge asymptotic behaviour must apply only to the smoothed out amplitude, and hence propose smearing the integrand on the l.h.s. to

$$\langle [A(s, t)]_s \rangle = (2\alpha(s) + 1) \bar{\beta}(s) P_{\alpha(s)}(1 + \frac{2t}{s - 4m^2}) \frac{d\alpha(s)}{ds} \quad (14.29)$$

The important point is that even for asymptotically large s , this has to hold for *every* t . It is highly unlikely that these sets of equations will yield one $\bar{\beta}(s)$. The best one can hope for is that as $s \rightarrow \infty$, $\bar{\beta}_t(s)$ approaches a single function for all values of t . Stated differently, $\bar{\beta}_t(s)$ has to be of the form $e^{\rho(s, t)} \bar{\beta}(s)$ with $\rho(s, t)$ *bounded*. Next they look at the FESR for very large s such that $s \gg t$. On using asymptotic form of Legendre functions for large arguments, one gets the condition

$$\bar{\beta}(s) e^{\rho(s, t)} I_0(2\alpha(s) \sqrt{\frac{t}{s}}) = \beta(t) s^{\alpha(t)} \quad (14.30)$$

This immediately shows that $\alpha(s)$ must rise *faster* than $\sqrt{s}$. Now t can also be taken to be very large, and upon using the asymptotic form of I_0 , one finally arrives at the condition

$$\bar{\beta}(s) e^{\rho(s, t)} e^{2\alpha(s) \sqrt{\frac{t}{s}}} = \beta(t) s^{\alpha(t)} \quad (14.31)$$

Given the boundedness of $\rho(s, t)$, no α can be found to satisfy this. This is the crux of their argument against finding a solution for FESR in the simple model.

Mandula and Slansky claim the arguments can be made much more general. In particular, they point out that additional trajectories do not make an essential difference. The essence of their argument is that the large s asymptotics is still governed by the leading trajectory on the right, and the same logic as above would dictate that must grow faster than $\sqrt{s}$. In effect, it becomes like a one trajectory model. Spin can change many details but again not the essence of the arguments.

So their final conclusion, in their own words is: *The Regge asymptotic behaviour, as expressed through FESR, is incompatible with any model in which scattering amplitudes are given by finitely spaced trajectories of direct-channel resonances.*

14.4.2 Fujisaki's Work

In yet another very incisive analysis [13], Haruo Fujisaki has exposed further fault-lines in the simple minded FESR and FESR-bootstrap with a few Regge trajectories. We state his conclusions first, and then go into their essential details. His conclusions are twofold: (a) In narrow resonance approximations, FESR can only be satisfied if all Regge trajectories (except the Pomeron) are accompanied by an *infinite* number

of *Daughter Trajectories* which are asymptotically parallel to the leading trajectories, and, (b) one cannot get effective constraints on asymptotic trajectory functions without additional assumptions.

Let us recollect that when direct-channel resonances are taken to lie on indefinitely rising Regge trajectories, FESR become consistency conditions on Regge trajectories in direct and crossed channels. At this stage, that can be taken as an intuitive characterization of Duality. An important ingredient here is the demonstration by van Hove [14] that an infinite number of resonances lying on an infinitely rising “tower” can be resummed as a Regge trajectory.

While we have clearly seen that FESR cannot be satisfied by a *single* trajectory in the direct channel, the question remains whether a *finite* number of such trajectories will suffice. According to Fujisaki the answer is not certain, but Mandula and Slansky did answer this in the negative.

Fujikawa analyses the process $\pi\pi \rightarrow \pi\omega$ with focus on the ρ -trajectory. He applies FESR to the helicity amplitudes. A matter of technicality is that they should be free of *Kinematic Singularities*. His inputs are Regge Asymptotic behaviour, crossing symmetry, narrow resonance approximation, as well as *Real Analyticity* of Regge parameters. On the basis of these he arrives at his chief conclusions, already stated above, for the ρ -trajectory. He acknowledges that similar results had also been obtained by Swift and Tucker [15].

The main advantages to considering $\pi\pi \rightarrow \pi\omega$ are, (i) only one family of Regge trajectories with the quantum numbers $I^{GC} = 1^{+-}$, $P = -$, and, (ii) only one independent amplitude which can be taken as the t -channel helicity amplitude $F_{\lambda 0}^t(s, t)$, $\lambda = \pm 1$. This amplitude is free of kinematic singularities in *all* channels and has the correct s, u symmetry.

The leading ρ -trajectory is accompanied by an infinite number of *Daughters* spaced by *two* units of angular momentum at $t = 0$, i.e. $\alpha_{D,n}(0) = \alpha_\rho(0) - 2n$. The details of $\alpha_{D,n}(t)$ depend on dynamics. Asymptotically, they are all *parallel* to the leading trajectory. With his inputs (see above), Fujisaki determines $\beta_{10}^i(t)$ to be of the form:

$$\beta_{10}^i(t) = G(\alpha_i(t)) \tilde{\beta}_i(t) \quad (14.32)$$

where G is some known function and $\tilde{\beta}$ an entire function.

Fujisaki also encounters the same situation encountered by Mandula and Slansky, i.e. of the r.h.s of FESR being a smooth function of $\bar{\nu}_N$ where ν_N is the energy variable and N playing the same role as the one in the case of Horn and Schmid, while the l.h.s is a sum of step functions. He too resorts to smearing. We leave out a discussion of all the details and simply quote his final results: A finite number of Daughter trajectories is insufficient to satisfy FESR and when the number of Daughters is infinite, FESR alone is not sufficient for Bootstrap.

Frampton too makes similar statements in his book [1] (see p.25 last few lines) regarding this. To quote him, “Not surprisingly, it can be shown that to find an analytic solution of FESR valid for all t , an infinite number of resonances and correspondingly an infinite number of Regge poles is necessary”.

14.5 Concluding Remarks

We have taken the trouble to give an in-depth discussion of superconvergence relations as well as of FESR, because of the very central roles they play in the major narrative of this book. The sections in this chapter exposing the lacunae in the solutions of the FESR conditions were also done in somewhat elaborate detail in anticipation of the path-breaking work of Veneziano [16] (to be taken up in the next chapter). That work, apart from making the notion of Duality mathematically precise, is also a beautiful illustration of how all these lacunae get resolved in an almost miraculous way.

Despite all these serious theoretical difficulties, people continue to use FESR in a somewhat practical sense to relate low energy and high energy parametrizations. A good example is the rather recent (2017) analysis by Nys [17], which is very much in the same spirit as one of the earliest phenomenological applications of FESR by Ademollo and collaborators [10].

References

1. P.H. Frampton, *Dual Resonance Models* (Benjamin, W.A, 1974)
2. S. Gasiorowicz, *Elementary Particle Physics* (John Wiley and Sons)
3. V. de Alfaro, S. Fubini, G. Rossetti, G. Furlan, Phys. Letts. **21**, 576 (1966)
4. K. Igi, Phys. Rev. Lett. **9**, 76 (1962)
5. D. Horn, C. Schmid, Caltech Preprint CALT-68-127
6. R. Dolen, D. Horn, C. Schmid, Phys. Rev. **166**, 1768 (1968)
7. A. Logunov, L.D. Soloviev, A.N. Tavkhelidze, Phys. Letts. B. **24**, 181 (1967)
8. S. Mandelstam, Phys. Rev. **166**, 1768 (1968)
9. M. Ademollo, H.R. Rubinstein, G. Veneziano, M. Virasoro, Phys.Rev.Lett. **19**, 1402 (1967); Phys. Lett. B. **27**, 99 (1968)
10. M. Ademollo, H.R. Rubinstein, G. Veneziano, M. Virasoro, Phys. Rev. **176**, 1904 (1968)
11. K. Igi, S. Matsuda, Phys. Rev. Lett. **18**, 625 (1967)
12. J.E. Mandula, R.C. Slansky, Phys. Rev. Lett. **20**, 1402 (1968)
13. H. Fujisaki, Prog. Theor. Phys. **43**, 101 (1970)
14. L. van Hove, Phys. Lett. B. **24**, 183 (1967)
15. A.R. Swift, R. Tucker, Phys. Rev. Lett. **22**, 1411 (1968)
16. G. Veneziano, Nuovo Cim. A. **57**, 190 (1968)
17. J. Nys, *Finite Energy Sum Rules: Going High to Solve low-energy Issues*, in NSTAR 2017. Columbia (2017)

The Veneziano Formula and the Dual Resonance Model

15

15.1 Introduction

This chapter describes the *watershed* moment in the entire development of the *The Analytic S-matrix*, which also more or less directly led to *Hadronic String Theory*, which then led to the *The Superstring Theory*. That moment was the paper in 1968 by Veneziano [1]. Titled *Construction of a Crossing-Symmetric, Regge-Behaved Amplitude for Linearly Rising Trajectories*, it gave an unexpectedly simple and lucid formula for the elastic scattering of two particles. What was remarkable about this formula was that despite its simplicity, it had achieved many important goals of the analytic S-matrix programme: Regge asymptotics, crossing symmetry, satisfying superconvergence relations and finite energy sum rules, duality of Regge poles and resonances, how analyticity can relate masses and many more.

It was in preparation for this that we discussed all these topics threadbare, both technically and conceptually, to enable the reader to truly appreciate the *Synthesis* brought forth by Veneziano's paper. In fact, the evolution of ideas that led to superconvergence relations, FESR, Bootstrap and finally the concept of Duality culminated naturally in Veneziano's work. In all the earlier discussions, a precise characterization of what Duality was remained elusive. The beauty of Veneziano's formula is that it demonstrated, in a mathematically precise way, what Duality means. The same amplitude can be expressed as due to exchange of t -channel trajectories or s -channel trajectories.

The formula also vindicated the critiques of the FESR and FESR-bootstrap as elaborated in the works of Mandula and Slansky on the one hand [2], and, in Fujikawa's work on the inevitability of *infinitely many Daughters* on the other [3] (this work actually appeared after Veneziano's paper). Both these features can be made manifest in Veneziano's formula.

This chapter also extends the discussion to 5-point functions, N-point functions, their factorizability, as well as the so-called *Operator Formalism*. It concludes with a

discussion of the *Dual Resonance Models* and their problems. A thorough discussion of the Koba-Nielsen variables is also given.

15.2 The Veneziano Formula

We first write down Veneziano's magic formula, and then discuss various properties of the formula, as well as the beautifully intuitive way Veneziano arrived at it. In discussing the many properties of the formula, we shall make use of the lucid analysis of Veneziano himself, along with the detailed exposition of the formula by Frampton in his book [4].

Veneziano's approach is based on the requirements of Regge asymptotics, crossing symmetry, and, linearly rising Regge trajectories, as envisaged by [5–7]. Veneziano considers the process $\pi\pi \rightarrow \pi\omega$. The pions are scalars while ω is a spin-one particle. This was the same process that Fujisaki had considered in [3] because of its many nice features. Among them is the fact that it is the same process in all the channels, i.e. s , t , u -channels. Crossing symmetry would therefore require *total symmetry* among the s , t , u variables as far as the *Invariant amplitude* is concerned which is defined by Veneziano to be

$$T = \epsilon_{\mu\nu\rho\sigma} e_\mu p_{1\nu} p_{2\rho} p_{3\sigma} A(s, t, u) \quad (15.1)$$

with T being the T -matrix element, e_μ the polarization of the *Vector Meson* ω , and, p_i the four-momenta of the three pions. The amplitude $A(s, t, u)$ is free of *kinematic singularities*. The formula for $A(s, t, u)$ that Veneziano proposed is

$$A(s, t, u) = \frac{\bar{\beta}}{\pi} \{B(1 - \alpha(t), 1 - \alpha(s)) + B(1 - \alpha(t), 1 - \alpha(u)) + B(1 - \alpha(s), 1 - \alpha(u))\} \quad (15.2)$$

where $\bar{\beta}$ is some *constant*, and, B is the well known Euler B-function:

$$B(x, y) = \frac{\Gamma(x) \Gamma(y)}{\Gamma(x + y)} \quad (15.3)$$

with $\Gamma(x)$ the Euler Gamma function with well established properties that will be crucial in understanding the true power of the Veneziano formula.

15.2.1 Veneziano's Motivation

Veneziano's analysis begins with a parametrization for $A(s, t, u)$ for asymptotically large s at fixed t , discussed in [8]:

$$A(s, t, u) \simeq \frac{\bar{\beta}}{\pi} \Gamma(1 - \alpha(t)) (-\alpha(s))^{\alpha(t)-1} + (s \leftrightarrow u) \quad (15.4)$$

which was in reasonable agreement with superconvergence relations. The main problem with this was that the crossing symmetry, which requires full symmetry in all of the three (s, t, u) variables is absent. It cannot be symmetrized by brute force as (s, t) symmetry would introduce unwanted poles (see Veneziano's paper for further details). Veneziano's approach to this difficulty was different: he sought to replace the $(-\alpha(s))^{\alpha(t)-1}$ factor, introduced for reasons of correct Regge asymptotics, by another expression with the same asymptotic s behaviour, but was more amenable to imposing crossing symmetry! Indeed, he found such a replacement to be

$$(-\alpha(s))^{\alpha(t)-1} \rightarrow \frac{\Gamma(1 - \alpha(s))}{\Gamma(2 - \alpha(s) - \alpha(t))} \quad (15.5)$$

Various properties of the Gamma function are essential for this realization. An unexpected fringe benefit of this replacement is that at least in principle, the formula can be considered even when s is not asymptotically large because of the *subleading* terms inherent in the Gamma function. The reader is referred to Sect. 6 (there is a typo in the paper and this section also appears as Sect. 5) of his paper for further details. This also clarifies another important aspect of the Veneziano formula: though superficially it appears as if there is only *one* trajectory function α , appearing as it does as an argument of the Gamma function, effectively there are *infinitely* many *Daughter* trajectories. The full form of the Veneziano formula is essential for this interpretation. We shall return to a more elaborate discussion of this later on.

15.2.2 Important Properties of the Veneziano Formula-I

We list here many important properties of the formula as elaborated by Veneziano himself in the paper. We will take up some of these later for a deeper discussion:

- **Regge Asymptotics:** Veneziano, in Sect. 1.0, shows how his formula reproduces the large s , fixed t asymptotic behaviour for the amplitude which Ademollo et al [8] had argued for:

$$A(s, t, u)_{Ven} \rightarrow \bar{\beta}(t) \frac{1 - \cos \pi \alpha(t)}{\Gamma(\alpha(t) \cdot \sin \pi \alpha(t))} [\alpha(s)]^{\alpha(t)-1} \quad (15.6)$$

- **Case of $\text{Im } \alpha \neq 0$:** As per one of the footnotes on the title page, Veneziano says that the paper mostly works within the approximation of real, linear trajectories. Strict reality of the trajectory function is only compatible with resonances of zero width. This is clearly unphysical and violates *unitarity*. Likewise, strictly linear trajectory functions also violate unitarity. However, there can be regimes where unitarity corrections are small, and the above-mentioned approximations realistic. Eventually trajectory functions must turn complex and the linear rise should also stop. Towards the end of Sect. 1, Veneziano makes some comments about the situations where $\text{Im } \alpha$ is non-vanishing. He remarks that in those general cases,

it is likely that the complex J -plane will have more complicated singularities like cuts, in addition to the moving Regge poles.

- In Sect. 4, he discusses in detail how his formula satisfies the superconvergence sum rules. He too has to *smoothen* the form of $\text{Im } A(s, t, u)$ which on its own would be a sum of delta functions. This aspect was already discussed in Chap. 14. So, the Veneziano formula is an explicit counter to various lacunae about satisfying FESR, as brought out in the works of Mandula and Slansky, and, Fujisaki. We will not go into the details of Veneziano's calculations.
- In Sect. 7, Veneziano applies his formula to derive the *fixed-angle* scattering when all of (s, t, u) approach very large values. He finds that his formula can account for *secondary dips and peaks*; in the past, this was a contentious issue which some people had attempted to solve by adding additional trajectories, but as had been pointed out by many, their parameters were not consistent with FESR.
- Veneziano also shows how his formula gives a *precise* characterization of Duality in the sense that the same amplitude can be thought as due to either s -channel trajectories, or, t -channel trajectories, but not both. Equally important, his formula does not allow double poles because of the $\Gamma(2 - \alpha(s) - \alpha(t))$ factor in the *denominator*. In his Sect. 5, he also shows how the formula correctly incorporates the dual aspects of resonances and Regge poles (not to be confused with Duality between direct and cross-channel trajectories). We shall present a more technical elaboration of these important aspects later where we shall make use of Frampton's analysis [4].
- **Issue of many trajectories:** Modulo the issue of *Daughter Trajectories* the Veneziano formula gives the impression of having used only *one* trajectory $\alpha(x)$, and whether that is an unnecessary restriction. The formula was proposed only for the invariant amplitude for $\pi\pi \rightarrow \pi\omega$ where crossing symmetry demanded complete symmetry in all of (s, t, u) . Veneziano clarifies this with another example where he considers the process $\pi\eta \rightarrow \pi\rho$. Crossing symmetry only requires a $s \leftrightarrow u$ symmetry. The quantum numbers are such that *two* trajectories, namely, the A_2, ρ , are required. The corresponding Veneziano formula takes the form:

$$A_{\pi\eta \rightarrow \pi\rho} = \frac{\bar{\beta}_1}{\pi} \{B(1 - \alpha_{A_2}(s), 1 - \alpha_\rho(t)) + B(1 - \alpha_{A_2}(u), 1 - \alpha_\rho(t)) - B(1 - \alpha_{A_2}(s), 1 - \alpha_{A_2}(u))\} \quad (15.7)$$

Here too, the infinitely many Daughters of both the A_2, ρ -trajectories contribute.

- **Relations among the masses:** Veneziano points out that his formula for $\pi\pi \rightarrow \pi\omega$ admits poles at *even* values of α which are not permitted by the angular momentum conservation laws. On noting that the residue of the pole at, say $\alpha(t) = 2$, is proportional to $\alpha(s) + \alpha(u)$, he proposes to eliminate the unwanted poles by demanding $\alpha(s) + \alpha(u) = 0$. Combining with $\alpha(t) = 2$, he then proceeds to propose the manifestly crossing-symmetric condition:

$$\alpha(s) + \alpha(t) + \alpha(u) = 2 \quad (15.8)$$

On combining this with the self-consistency requirement $\alpha(m_\rho^2) = 1$ (i.e. the ρ -resonance must lie on the ρ -trajectory) along with the assumption of linearly rising Trajectories, i.e. $\alpha(x) = \alpha(0) + \alpha' x$, Veneziano immediately obtains

$$\alpha(-2m_\rho^2 + m_\omega^2 + 3m_\pi^2) = 0 \quad (15.9)$$

Such a relation had been obtained earlier from *sum rules*. It can be rearranged as

$$m_\rho^2 = \frac{m_\omega^2}{3} + m_\pi^2 + \frac{1}{3\alpha'_\rho} \quad (15.10)$$

and can be thought of as a relation between masses ($\frac{1}{\alpha'_\rho}$ having dimensions of M^2 can also be interpreted as one of the masses). In Chap. 14, we had seen how super-convergence relations related masses and coupling constants. In contrast, what is obtained now is a relationship between only masses. It is also worth pointing out that such a relationship between masses is not obtainable from internal symmetry arguments like SU(3) as the relation mixes (pseudo) scalar and vector meson masses.

Veneziano makes a similar analysis for $\pi\eta \rightarrow \pi\rho$; to remove unwanted poles he demands

$$\alpha_{A_2}(s) + \alpha_{A_2}(u) + \alpha_\rho(t) = 2 \quad (15.11)$$

Note that this relation too is crossing-symmetric ($s \leftrightarrow u$, in this case). For linearly rising trajectories, this immediately requires $\alpha'_{A_2} = \alpha'_\rho = \alpha'$ (as the above equation has to hold for all s, t, u subject to the kinematic constraint $s + t + u = 2m_\pi^2 + m_\eta^2 + m_\rho^2$). Then it easily follows, on using $\alpha_\rho(m_\rho^2) = 1$ that

$$\alpha' = \frac{1}{3m_\rho^2 - m_\omega^2 - 3m_\pi^2} \quad \alpha_{A_2}(0) = 1 - \frac{1}{2} \frac{3m_\rho^2 - m_\omega^2 - m_\pi^2 + m_\eta^2}{3m_\rho^2 - m_\omega^2 - 3m_\pi^2} \quad (15.12)$$

One can go a step further by combining these to give the explicit formula for $m_{A_2}^2$:

$$m_{A_2}^2 = \frac{3m_\rho^2 - m_\omega^2 - m_\pi^2 + m_\eta^2}{2} \quad (15.13)$$

This is a genuine mass relation. Again it mixes scalar and vector meson masses! From this Veneziano estimates $m_{A_2} \simeq 1350\text{MeV}$ while the observed value is approximately 1320 MeV!

15.2.3 Precise Duality

So far, we had not given a very clear idea of exactly what Duality meant. It's meaning was rather implicit. It roughly meant that the FESR and superconvergence relations implied that there was a connection between direct-channel Regge poles and crossed-channel Regge poles; that it is enough to construct the scattering amplitude with one, or the other, but not both. In a certain vague sense, it also meant that the amplitudes did not have *double poles*, one in each channel. Now we will see how the Veneziano formula encapsulates the important idea of *precise duality*.

We shall follow our earlier approach of using the simplest of scattering phenomena to illustrate the basic principles. From that point of view it is better to start with a Veneziano-like formula for the two-body elastic scattering of two neutral, scalar particles. This is indeed what Frampton does in [4]. In contrast, in both of the examples discussed by Veneziano, there are spin-1 particles involved, ω in the first, and, ρ in the second. The Veneziano formula for the case of all neutral scalar particles becomes (see [4] Sect. 2.2 Eq. (2.1)):

$$A_4(s, t, u) = \bar{\beta} \{ B(-\alpha(s), -\alpha(t)) + B(-\alpha(t), -\alpha(u)) + B(-\alpha(u), -\alpha(s)) \} \quad (15.14)$$

with B the same Euler B-function introduced earlier. Here too, A_4 is completely symmetric in (s, t, u) as demanded by crossing symmetry, and as can be seen from the fact that $B(x, y) = B(y, x)$.

The Gamma function $\Gamma(z)$ has *simple poles* in z whenever z is either zero or a negative integer. The absence of double poles can be immediately seen in terms of the representation of Eq. (15.3). For double poles to occur in $B(x, y)$, the numerator would require *both* (x, y) to take these special values, but then so will $x + y$ and the $\Gamma(x + y)$ in the denominator would kill the double pole!

To see the more detailed aspects of the poles and their residues of $B(x, y)$, one could have used the well-known properties of $\Gamma(z)$ itself. But as Frampton has pointed out (see his Eq. (2.7) [4])

$$\Gamma(z) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{1}{z+n} + \Gamma(1, z) \quad (15.15)$$

here $\Gamma(1, z)$ is the so-called *Incomplete Gamma function*. It is an analytic function. Because of this additional piece, it becomes clumsy to expose the *precise duality* alluded to before. Instead, it is more straightforward to use the *integral representation* for $B(x, y)$ [4]:

$$B(x, y) = \int_0^1 dt t^{x-1} (1-t)^{y-1} \quad (15.16)$$

As we shall see, such integral representations not only throw light on the origin of precise duality, they are also very important for generalizing the Veneziano amplitude

for multiparticle processes. Since the range of integration is 0 to 1, it is legitimate to Taylor expand $(1-t)^{y-1}$ around $t=0$:

$$(1-t)^{y-1} = \sum_{n=0}^{\infty} \frac{(-t)^n}{n!} \frac{\Gamma(y)}{\Gamma(-n+y)} \quad (15.17)$$

This is nothing but the usual Binomial expansion, now generalized to non-integral powers. It is then easy to show that

$$B(x, y) = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{1}{x+n} (1-y)(2-y) \dots (n-y) \quad (15.18)$$

It should be noted that the order of summation and integration has been interchanged, which is justified here. Thus the poles and residues structure of $B(x, y)$ is simple indeed.

In this form the symmetry $B(x, y) = B(y, x)$ is not manifest. By changing the variable of integration above from $t \rightarrow 1-t$, the integral representation can be made to look like

$$B(x, y) = \int_0^1 dt t^{y-1} (1-t)^{x-1} \quad (15.19)$$

Evaluating this by Taylor expanding $(1-t)^{x-1}$ yields

$$B(x, y) = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{1}{y+n} (1-x)(2-x) \dots (n-x) \quad (15.20)$$

These two can be combined into what is a precise statement of Duality [4]

$$B(-\alpha(s), -\alpha(t)) = \sum_{n=0}^{\infty} \frac{R_n(t)}{n - \alpha(s)} = \sum_{n=0}^{\infty} \frac{R_n(s)}{n - \alpha(t)} \quad (15.21)$$

with

$$R_n(x) = \frac{1}{n!} (\alpha(x) + 1)(\alpha(x) + 2) \dots (\alpha(x) + n) \quad (15.22)$$

This realizes Duality in a *mathematically precise* way in the sense that the amplitude can either be thought of as entirely due to the s -channel (direct channel) or entirely due to t -channel (crossed channel) Regge trajectories, but not both at the same time. The absence of the *double poles* is obvious in the above. The numerator functions R_n are polynomials of degree n in the variable t for the s -channel pole and vice versa. On using the relations between t , u , and the scattering angle θ_s , more precisely $z_s = \cos \theta_s$

$$t = \frac{s - 4\mu^2}{2} (z_s - 1) \quad u = -\frac{s - 4\mu^2}{2} (z_s + 1) \quad (15.23)$$

it is seen that $R_n(t)$, $R_n(u)$ both consists of *spin-exchanges* with maximum spin being, and other spins $n-1, n-2, \dots, 0$. The largest spin of n is referred to as the *Parent* and all the lower ones as *Daughters*.

There is a point that has not been emphasized sufficiently in [4]; it is that the numerator functions $R_n(t)$, $R_n(u)$ still have a dependence on s . It is only when they are evaluated at the location of the pole, say in s , at $s_n = \frac{n-\alpha(0)}{\alpha'}$ that they become true residues. Making this distinction explicit

$$\bar{R}_n(t) = \xrightarrow{s \rightarrow s_n} R_n(t) \quad \bar{R}_n(u) = \xrightarrow{s \rightarrow s_n} R_n(u) \quad (15.24)$$

In this sense, $R_n(t)$ contain terms proportional to $(n - \alpha(s))$, $(n - \alpha(s))^2 \dots$ Likewise for $R_n(u)$, and these contribute to regular terms. These are undoubtedly related to $\Gamma(1, z)$ we encountered before. Precise duality relates these regular terms in the direct and the cross-channels. In what follows we shall only use the residue functions $\bar{R}_n(t)$, $\bar{R}_n(u)$.

It is obvious that both $R_n(x)$ and $\bar{R}_n(x)$ are, for linearly rising trajectories, of the form

$$R_n(x) = (\alpha' x)^n + c_1(\alpha(0)) (\alpha' x)^{(n-1)} + \dots + c_n(\alpha(0)) \quad (15.25)$$

The full scattering amplitude $A_4(s, t, u)$ expressed in terms of the s-channel poles takes the form

$$A_4(s, t, u) = \bar{\beta} \sum_{n=0}^{\infty} \frac{1}{n - \alpha(s)} (R_n(t) + R_n(u)) \quad (15.26)$$

Of course, due to the precise duality, this could have been expressed as a sum over t (or u)-channel poles with the corresponding R_n -functions. Now

$$R_n(t) + R_n(u) = \sum_{r=0}^n c_r(\alpha(0)) (\alpha')^{n-r} (t^{n-r} + u^{n-r}) \quad c_0 = 1 \quad (15.27)$$

Now for any positive integer m (both even and odd)

$$t^m + u^m = \left(\frac{s_n - 4\mu^2}{2}\right)^m ((z_s - 1)^m + (-1)^m (z_s + 1)^m) \equiv \mathcal{P}_{m,n}(z_s) \quad (15.28)$$

It is clear that the polynomials $\mathcal{P}_{m,n}$ satisfy

$$\mathcal{P}_{m,n}(z_s) = \mathcal{P}_{m,n}(-z_s) \quad (15.29)$$

In other words, the residues of the poles for *all* values of n contain only *even* spins. This is as required by Bose statistics as two identical spin-less particles can only couple to even spin particles.

It should be noted that this result is *generic* and holds irrespective of the values of $\alpha(0)$, α' .

15.2.4 The Issue of Daughters

The need for Daughter trajectories with intercepts separated from that of the Parent trajectory by *even* integers had been noted even before Veneziano's classic work [8,9]. The importance of Daughter trajectories for consistent solutions to FESR was also discussed at length in [3] though this work appeared after Veneziano's paper. According to these works, only *even* Daughters separated by the Parent trajectory by $2n$ units should be present.

In Veneziano's original work, where processes like $\pi\pi \rightarrow \pi\omega$ and $\pi\eta \rightarrow \pi\rho$ were concerned, this amounted to the requirement that poles at $\alpha(t) = 2m$ should be absent. He explicitly found that requiring the absence at $\alpha(t) = 2$ miraculously guaranteed the absence of the poles at all even values!

We now repeat that analysis for the spin-less particle scattering under consideration. Now, it is the poles at odd values, i.e. $\alpha(s) = 2n + 1$ (Veneziano considered the t -channel version of this) that should be absent. This requirement should be clearly distinguished from the absence of odd-spins in both the odd and even residue functions that were already demonstrated.

Let us first consider the case of $\alpha(s) = 1$. The corresponding residue for the full amplitude is $(1 + \alpha(t)) + (1 + \alpha(u))$ which must vanish. Just as Veneziano did, the two conditions can be combined into one

$$2 + \alpha(t) + \alpha(u) + \alpha(s) = 1 \quad (15.30)$$

For linear trajectories this translates to

$$3\alpha(0) + 4\alpha'\mu^2 = -1 \quad (15.31)$$

where, as before, μ is the common mass of the four particles. At this stage, both Veneziano [1] and Frampton [4] invoke the so-called *Bootstrap Condition* requiring the scattering particles to also lie on the trajectory $\alpha(s)$, which in the present context implies $\alpha(\mu^2) = 0$ which further implies

$$\alpha(0) + \alpha'\mu^2 = 0 \quad (15.32)$$

On combining the last two conditions, one obtains $\alpha(0) = 1$ as the requirement for the first odd-daughter to be absent. For a linearly *rising* trajectory, i.e. $\alpha' > 0$, this immediately implies that $\mu^2 < 0$! In other words the scattering particles are *Tachyonic*! This violates the standard *Spectrum Conditions* of Quantum Field Theories and is usually considered undesirable except perhaps by George Sudarshan! We will soon see that this feature is unavoidable in a class of String Theories to follow.

Frampton shows in detail that the consequence of $\alpha(0) = 1$ is the property of the residue functions

$$\bar{R}_n(t) = (-1)^n \bar{R}_n(u) \quad (15.33)$$

which immediately ensures the absence of all odd-daughters, leaving only the even-daughters which are spaced $2n$ units from the Parent α . Veneziano also makes brief

comments on the analogous absence of even-daughters in his examples, again leading to daughters that are also spaced $2n$ units from the corresponding Parent α_ρ .

This is indeed rather miraculous that just requiring the absence of the first daughter automatically ensures the removal of all the unwanted daughter trajectories. The impression one gathers from the treatments of both Veneziano and Frampton is that somehow the *Bootstrap Condition* is essential for this. As this concerns an important issue of principle, we shall see how far one can go *without* invoking the Bootstrap condition, i.e. with only Eq. (15.31). To our surprise, we find that this alone is sufficient to guarantee the absence of all odd-daughters!

Let us start with the n th s -channel pole $\alpha(s_n) = n$. It is easy to see that for a linear trajectory $s_n = \frac{n - \alpha(0)}{\alpha'}$. To analyse the properties of the residue functions $\bar{R}_n(t)$, $\bar{R}_n(u)$ we first recall $t = \frac{s - 4\mu^2}{2}(z_s - 1)$, $u = -\frac{s - 4\mu^2}{2}(z_s + 1)$ where z_s is the s -channel scattering angle. Consequently, the interchange $t \leftrightarrow u$ is equivalent to $z_s \leftrightarrow -z_s$. The trick is to find a m such that

$$m + \alpha(0) = \alpha' \frac{s_n - 4\mu^2}{2} \quad (15.34)$$

The solution on using the condition Eq. (15.31) and with no other conditions is easy to find:

$$m = \frac{n + 1}{2} \quad (15.35)$$

We now analyse the odd and even residue functions $\bar{R}_{2n+1}(x)$, $\bar{R}_{2n}(x)$ separately. Let us consider the odd case $2n + 1$ first. Therefore, $m = n + 1$, and, Eq. (15.34) becomes $n + 1 + \alpha(0) = \alpha' \frac{s_{2n+1} - 4\mu^2}{2}$. In $\bar{R}_{2n+1}$ let us focus on the factor $(n + 1 + \alpha(t))$; it is easy to show that it equals $(n + 1 + \alpha(0))z_s$! The two factors adjacent to this central term are $(n + \alpha(t))$ and $(n + 2 + \alpha(t))$; they are seen to equal $(n + 1 + \alpha(0))z_s - 1$ and $(n + 1 + \alpha(0))z_s + 1$ respectively, with their product equal to $(n + 1 + \alpha(0))^2 z_s^2 - 1$. Likewise, the product of the first and the last terms is seen to be $(n + 1 + \alpha(0))^2 z_s^2 - n^2$. Consequently, on introducing $\xi = (n + 1 + \alpha(0))z_s$,

$$\bar{R}_{2n+1}(t) = \xi \cdot \prod_{p=1}^n (\xi^2 - p^2) \quad (15.36)$$

Recalling that $t \leftrightarrow u$ is now equivalent to $\xi \leftrightarrow -\xi$, it is easily seen that

$$\bar{R}_{2n+1}(t) = -\bar{R}_{2n+1}(u) \quad (15.37)$$

which is precisely the condition for the vanishing of the residue at odd poles for the whole amplitude. And we have arrived at this *without* invoking the Bootstrap Condition! There is one more requirement and that is $\bar{R}_{2n}(t) = \bar{R}_{2n}(u)$. To show this, we group the factors in $\bar{R}_{2n}$ as $(n, n + 1)$, $(n - 1, n + 2) \dots (1, 2n)$. The $(n, n + 1)$ pair can be worked out, after using the fact that now $m = n + \frac{1}{2}$,

to be $(n + \frac{1}{2} + \alpha(0))z_s - \frac{1}{2}$ and $(n + \frac{1}{2} + \alpha(0))z_s + \frac{1}{2}$ and their product to be $(n + \frac{1}{2} + \alpha(0))^2 z_s^2 - (\frac{1}{2})^2$. Now introducing $\eta = (n + \frac{1}{2} + \alpha(0))z_s$, the result is

$$\bar{R}_{2n}(t) = \prod_{p=1}^n (\eta^2 - (p - \frac{1}{2})^2) \quad (15.38)$$

The $t \leftrightarrow u$ being now equivalent to $\eta \leftrightarrow -\eta$, it follows that $\bar{R}_{2n}(t) = \bar{R}_{2n}(u)$, again without invoking the Bootstrap Condition!

Huang [10] has given an even cleverer way of arriving at these. Though his original argument was given for Veneziano's original amplitude, it can be easily adapted to the present discussion of the neutral spin-less scattering case. His reasoning is that the condition $\alpha(s) + \alpha(t) + \alpha(u) = -1$ for the absence of the first daughter implies, for the odd-pole $\alpha(s) = 2n + 1$, that $(1 + \alpha(t)) = -(2n + 1 + \alpha(u))$. This immediately guarantees $R_{2n+1}(t) = -R_{2n+1}(u)$. For the even residues, an identical argument works. This too shows that the Bootstrap condition is not required in the proofs.

It is therefore clear that the proof of the absence of all odd-daughters given that the first of them is absent *does not* require the Bootstrap Condition. If that condition is invoked it immediately leads to $\alpha(0) = 1$ with all the subsequent consequences, of which we have already mentioned the presence of Tachyons in the spectrum.

When $\alpha(0) = 1$, it is easy to see that

$$\bar{R}_2(t) = \bar{R}_2(u) = \frac{25}{16} (z_s^2 - \frac{1}{25}) \quad (15.39)$$

On comparing with the analogues of Legendre Polynomials in higher dimensions, the so-called *Hyperspherical Legendre Polynomials* $P_{n,D}(z)$ for the n th polynomial in D space-time dimensions (for an extensive treatment see Compos and Cunha [11]), for $n = 2$,

$$P_{2,D}(z) = \frac{(D-1)(D-3)}{2} (z^2 - \frac{1}{D-1}) \quad (15.40)$$

we see that R_2 contains only spin-2 in exactly $D = 26$ dimensions! This particular value of D will continue to play a very special role in all our future discussions.

So what exactly are we to make of the Bootstrap Condition? What if the scattered particles are spin-less particles lying on one of the many Daughter trajectories? Will the form of the Veneziano amplitude change for them? They too will have their own variants of the absence of odd-daughters. Will that be compatible with what has already been found? If the particles are spin-less but not *identical*, as would be the case when they would lie on different daughter trajectories? We leave it to the reader to have fun exploring them!

15.3 The Multi-point Function Generalizations

It must be fairly evident from our elaborate discussions of LSZ reductions, analyticity, dispersion relations, etc. that moving from $2 \rightarrow 2$ scattering to *production* processes like $2 \rightarrow n$; $n \geq 3$ will be technically formidable! In that sense, Veneziano's work opened up novel windows to these important hadronic processes. Of course, they would have the same limitations as Veneziano's 4-point function, namely, narrow resonance approximation, linearly rising trajectories, etc., but just as the four-point function formula had a number of conceptual as well as phenomenological successes, these multi-point functions will also have those positive features.

15.3.1 The 5-Point Function

In this subsection, we shall take a look at the simplest production process described by a *Five-Point Function*. The analysis is facilitated by taking all five momenta to be *ingoing*, irrespective of whether the particles with positive energy are ingoing or outgoing. The sign of the energy will keep track of it. The overall energy-momentum conservation condition reads

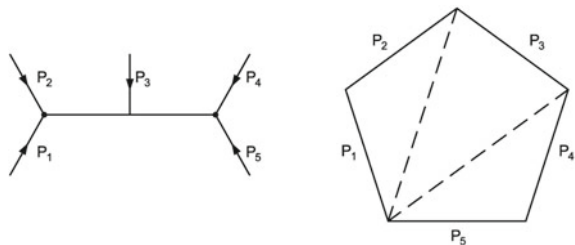
$$p_1 + p_2 + p_3 + p_4 + p_5 = 0 \quad (15.41)$$

There are five *mass invariants* $p_i \cdot p_i$; $i = 1, \dots, 5$. Whether the particles are actually on their mass shell or not is not relevant at this stage. If they are on the mass shell $p_i \cdot p_i = \mu_i^2$. In addition, there are 10 non-mass invariants $p_i \cdot p_j$, $i < j$. Crossing symmetry mixes these ten invariants. Because of the overall energy-momentum conservation of Eq. (15.41), only 5 of these are *independent* just as only two of the (s, t, u) invariants were independent for the four-point function. Once a choice of the five independent invariants has been made, it is possible to relabel the particles such that the invariants s_i correspond to the square of the total momentum of two *adjacent* particles. The crossing symmetry now takes the simpler form of $s_i \rightarrow s_{\pm i \pm 1}$. In the case of the four-point function, the total amplitude had to be symmetrized among all the three, i.e. (s, t, u) invariants; here too, after obtaining the analog of the Euler B-function of any of the five independent invariants, a final sum over inequivalent(modulo cyclic) permutations has to be performed. We shall not go into those specifics.

Immediately after Veneziano's work, Bardakci and Ruegg [12] and Virasoro [13] were among the first who generalized his formula to the five-point function. In addition to the case of neutral spin-less particles, Bardakci and Ruegg also generalized to spin-less particles with both isospin and strangeness [14]. It is remarkable that many important developments like these, culminating in the generalizations to the N-point function, all occurred within a short span of a few months towards the end of 1968 and beginning of 1969.

We shall base our discussion of the 5-point function on Virasoro's short (just about two pages!) and transparent paper as we find it more accessible pedagogically

Fig. 15.1 Channels allowed by duality. The figure on left is a multiperipheral diagram



but otherwise self-contained in all the technical aspects. We shall make comparisons with other treatments in [4] as well as in Nambu's pioneering work [15].

The first task is to distil the essential aspects (Duality) of the Veneziano 4-point function that will pave the way for the generalization to 5-point functions. These are, clearly, (i) the absence of double poles (say, in s and t), and, (ii) the residues of the poles in s are polynomials in $t(u)$. Now in the case of the 5-point function, there are 5 channels corresponding to each of the five s_i . In principle, three-body channels ought to be included too, but in the case of 5-point functions every 3-body channel can equally well be described as a two-body channel with the remaining two particles. Poles at s_i would have, as residues, functions (polynomials) of s_{i-1} and s_{i+1} . Therefore, double poles can only occur in *non-adjacent* channels (s_i, s_{i+2}), etc. Likewise, it follows that for the 5-point function, there can be no *triple* poles (see [13]).

There is a simple graphical algorithm to figure out all the possible channels and those channels that can have simultaneous poles. This is illustrated in Fig. 2.2 of Frampton's book [4]. For ease of reading, we show the essentials of it in Fig. 15.1.

For the 5-point function, one first draws a pentagon whose edges represent the 5-momenta in cyclic order (the edge lengths can be arbitrary). Then the possible channels, which are realized by two or more particles scattering to the others, are obtained by drawing all possible diagonals. Compatible channels will be those with *non-intersecting diagonals*. Quite obviously, all the diagonals drawn from any particular vertex will be compatible. For the pentagon, there are all together 5 diagonals out of which only two can be compatible channels. The absence of triple poles is reflected by the fact that only two diagonals can be non-intersecting. This geometric method works for any N -point function with a N -sided polygon playing the role of the pentagon. There is an equivalent method based on *multiperipheral diagrams*, but the polygon method is more straightforward. Both methods are illustrated in Frampton's book.

Both the 4 and the 5-point amplitudes are limited in the sense that there is only one *type* of the channel involved, $2 \rightarrow 2$ in the case of the 4-point function, and $2 \rightarrow 3$ in the 5-point case. As a convention, we shall label a channel for arbitrary N -point functions by (p, q) , $p \leq q$. This is to avoid double counting. In the 5-point example, $3 \rightarrow 2$ should not be counted separately. In that sense, 6-point functions are the first instances where more than one type of channel occurs viz. $2 \rightarrow 4$ and $3 \rightarrow 3$. Thus, 5-point functions may be inadequate in finding the N -point generalizations.

A quick look at the integral representation of the Veneziano four-point function is helpful in making further progress.

$$B(-\alpha(s), -\alpha(t)) = \int_0^1 dt t^{-\alpha(s)-1} (1-t)^{-\alpha(t)-1} \quad (15.42)$$

The singularities in $\alpha(s)$ come from the region $t \simeq 0$ of integration, while those in $\alpha(t)$ from the $t \simeq 1$ region, giving another explanation for the non-occurrence of double poles. Taylor expansions to all orders around $t = 0$ being completely equivalent to the analogous all-order Taylor expansions around $t = 1$ being the explanation for the amplitude built out of all poles in s -channel being completely equivalent to the same amplitude built out of all poles in the t -channel!

Based on these considerations, Virasoro starts with the following expression for the 5-point function B_5 :

$$B_5(s_1, s_2, s_3, s_4, s_5) = \int_R \prod_{i=1}^5 dx_i f_i(\{x_j\})^{-\alpha(s_i)-1} \quad (15.43)$$

in complete analogy with the integral representation for B_4 . R is the domain of integration in the 5-dimensional space to be determined. The absence of triple poles immediately implies that the relevant subspace of the 5-dimensional x -space is actually *two-dimensional* (see [13] for details):

$$B_5(s_1, s_2, s_3, s_4, s_5) = \int dx_1 dx_2 \prod_{i=1}^5 f_i(x_1, x_2)^{-\alpha(s_i)-1} \quad (15.44)$$

The above-mentioned restriction on double poles being permitted at best in non-adjacent channels means, if $f_i \simeq 0$ for some i , both f_{i-1} and f_{i+1} must be close to 1. These can be coded into the set of constraints

$$f_i = 1 - f_{i-1} f_{i+1} \quad (15.45)$$

These can be thought of as *Duality Constraints*. There are many solutions to these; one such is

$$f_1 = \frac{1-f_5}{1-f_3 f_5} \quad f_2 = \frac{1-f_3}{1-f_3 f_5} \quad f_4 = 1 - f_3 f_5 \quad (15.46)$$

Another such solution is

$$f_1 = \frac{1-f_2}{1-f_2 f_4} \quad f_5 = \frac{1-f_4}{1-f_2 f_4} \quad f_3 = 1 - f_2 f_4 \quad (15.47)$$

Not surprisingly, the two sets are related by the crossing map $f_i \rightarrow f_{i+1}$. With only two independent f_i , i.e. f_3, f_5 in the first case, and f_2, f_4 in the second, it is

legitimate to change the variables x_1, x_2 in Eq. (15.44) to, say, f_3, f_5 in the first case, and to f_2, f_4 in the second. In doing so, a non-trivial metric ρ in the f -coordinates has to be allowed. Had we changed the f -coordinates from (f_3, f_5) to (f_4, f_1) by crossing transformation, crossing symmetry would have required

$$\rho(f_3, f_5) df_3 df_5 = \rho(f_4, f_1) df_4 df_1 \quad (15.48)$$

Since $\frac{\partial(f_4, f_1)}{\partial(f_3, f_5)} = \frac{f_5}{f_4}$, any metric of the form

$$\rho(f_3, f_5) = \frac{\rho_I}{f_4} = \frac{\rho_I}{1 - f_3 f_5} \quad (15.49)$$

with ρ_I a *fully symmetric* function of all the s_i would ensure crossing symmetry. For simplicity, ρ_I can be chosen to be 1. This leads to Virasoro's integral representation for a dual 5-point function:

$$B_5(s_1, s_2, s_3, s_4, s_5) = \int \frac{df_3 df_5}{f_4} \prod_{i=1}^5 f_i^{-\alpha(s_i)-1} \quad (15.50)$$

which can be further rewritten as

$$B_5 = \int df_3 df_5 f_3^{-\alpha(s_3)-1} f_5^{-\alpha(s_5)-1} (1 - f_3)^{-\alpha(s_2)-1} \cdot (1 - f_5)^{-\alpha(s_1)-1} (1 - f_3 f_5)^{\alpha(s_1)+\alpha(s_2)-\alpha(s_4)} \quad (15.51)$$

We shall not analyse this amplitude further from the point of view of Regge asymptotics, precise duality, or structure of Daughters as was done for the 4-point function. We refer the reader to Virasoro's paper for those details. Our aim at this point is to understand the structure of the 5-point function in order to get to the N-point function. The 5-point function obtained by Bardakci and Ruegg (Eq. (6) of [12]) is

$$F_5^{BR} = \int du_1 du_4 u_1^{-\alpha_{12}-1} u_4^{-\alpha_{45}-1} \left(\frac{1 - u_1}{1 - u_1 u_4} \right)^{-\alpha_{23}-1} \cdot \left(\frac{1 - u_4}{1 - u_1 u_4} \right)^{-\alpha_{34}-1} (1 - u_1 u_4)^{-\alpha_{15}-2} \quad (15.52)$$

In the above, $\alpha_{ij} = \alpha(0) + \alpha'(p_i + p_{i+1} \dots + p_j)^2$. The last exponent has -2 in place of -1 because of the metric factor. The expression for the 5-point function obtained by Frampton in his Eq. (2.54) is (there are numerous typos both in this and the previous Eq. (2.52); after correcting for these, one obtains the next equation)

$$F_5^{Fra} = \int dx_1 dx_4 x_1^{-\alpha(s_1)-1} x_4^{-\alpha(s_4)-1} (1 - x_1)^{-\alpha(s_2)-1} \cdot (1 - x_4)^{-\alpha(s_3)-1} (1 - x_1 x_4)^{-\alpha(s_5)+\alpha(s_2)+\alpha(s_3)} \quad (15.53)$$

where now $s_i = (p_i + p_{i+1})^2$; $p_6 = p_1$. Finally, we make a comparison with the 5-point function that follows from the expression for the N-point function given by Nambu in [15] (we shall separately discuss the N-point functions in the next section):

$$\begin{aligned}
 B_n^{Nam} = & \int \frac{dx dy dz \dots dw}{(1-xy)(1-yz)\dots} x^{(12)} y^{(123)} z^{(1234)} \dots w^{(n-1,n)} \\
 & \cdot \left(\frac{1-x}{1-xy}\right)^{(23)} \left[\frac{(1-y)(1-xyz)}{(1-xy)(1-yz)}\right]^{(34)} \dots \\
 & \cdot \left(\frac{1-xy}{1-xyz}\right)^{(234)} \left[\frac{(1-yz)(1-xyzu)}{(1-xyz)(1-yzu)}\right]^{(345)} \dots
 \end{aligned} \quad (15.54)$$

Nambu's use of the alphabets $x, y, \dots, w$ for the channel variables does not lend itself to any easy generalizations. It is also confusing in the sense that it is not clear if the w variable should be present for every n or not. The other symbols like (ij) , (ijk) are defined by

$$\begin{aligned}
 (i, i+1) &\equiv -\alpha_2(s_{i,i+1}) - 1 = -(p_i + p_{i+1})^2 - a_2 - 1 \\
 (i, i+1, i+2) &\equiv -\alpha_3(s_{i,i+1,i+2}) - 1 = -(p_i + p_{i+1} + p_{i+2})^2 - a_3 - 1
 \end{aligned} \quad (15.55)$$

etc. The reason we have invoked Nambu's N-point function ahead of the other N-point proposals is to clarify a needless confusion in Frampton's discussion of the 5-point function, especially his Eq. (2.55). The confusion has to do with cases where the Regge intercepts can in principle be different. In Nambu's formula, the intercepts $a_2, a_3 \dots$ can in principle be different. Now in the case of the 5-point function, for reasons already explained there can only be one non-trivial intercept, and $a_3 = a_2$. Likewise, in this specific case $s_{123} = s_{45}$ and $s_{234} = s_{51}$. Then the Nambu 5-point function takes the form

$$\begin{aligned}
 B_5^{Nam} = & \int \frac{dx dy}{1-xy} x^{-\alpha(s_1)-1} y^{-\alpha(s_4)-1} \left(\frac{1-x}{1-xy}\right)^{-\alpha(s_2)-1} \\
 & \cdot \left(\frac{1-y}{1-xy}\right)^{-\alpha(s_3)-1} (1-xy)^{-\alpha(s_5)-1}
 \end{aligned} \quad (15.56)$$

It is straightforward to check that all the different 5-point functions are in fact the same, with only changes in parametrizations. This brings us to the alternate expression for the 5-point function given by Frampton in his Eq. (2.55):

$$\begin{aligned}
 B_5^{Fra2} = & \int dx_1 dx_4 x_1^{-\alpha(s_1)-1} x_4^{-\alpha(s_4)-1} (1-x_1)^{-2p_2 \cdot p_3 - c_2} \\
 & \cdot (1-x_4)^{-2p_3 \cdot p_4 - c_2} (1-x_1 x_4)^{-2p_2 \cdot p_4 - c_3}
 \end{aligned} \quad (15.57)$$

He defines c_n according to

$$c_n = \alpha_n(0) - 2\alpha_{n-1}(0) + \alpha_{n-2}(0) \quad (15.58)$$

To understand better what he is aiming at, let us look, for example, at the $(1 - x_1)$ term. The exponent of this term is $-\alpha(s_2) - 1$. Since $s_2 = (p_2 + p_3)^2 = 2\mu^2 + 2p_2 \cdot p_3$ and $\alpha(s) = \alpha(0) + \alpha' s$, it follows that α' has been set to $\alpha' = 1$. Here $p_i \cdot p_i = \mu^2$ and we do not yet assume the Bootstrap condition $\alpha(\mu^2) = 0$. Since α' is dimensionful, $\alpha' = 1$ means that all s_i are measured in units of $-\mu^2$ (as there is no other mass scale). It is clear that $\alpha_n(0)$ of Frampton correspond to a_n of Nambu for $n \geq 2$. What about $\alpha_1(0)$ and $\alpha_0(0)$? These do not correspond to any intercepts and Frampton, inspired by Nambu, defines them to be $\alpha_0(0) = 1$ and $\alpha_1(0) = -\alpha' \mu^2 = 1$.

Therefore, in the present context, $\alpha_2(0) = \alpha(0)$. Then Eq.(15.58) would give $c_2 = \alpha(0) - 1$ which differs from Frampton's Eq.(2.57) by a sign! It is crucial to correct this error in Frampton's treatment as we shall keep finding as we go along.

Now we turn to the $(1 - x_1 x_4)$ term. Its exponent is $-\alpha(s_5) + \alpha(s_2) + \alpha(s_3)$. To evaluate it we use a trick which will soon be seen to be the basis for a very important representation of all N-point functions, first introduced by Bardakci and Ruegg [21]. This form is crucial for realizing the so-called *Operator Formalism*, which in turn lays the foundation for the *String Theory* formulation. It is for this reason that we wish to go through these points very carefully.

The exponent is $\alpha(0) + \alpha'(-s_5 + s_2 + s_3)$. The trick is to rewrite the second part as $\alpha'(-(p_2 + p_3 + p_4)^2 + (p_2 + p_3)^2 + (p_3 + p_4)^2 - p_3 \cdot p_3) - 1$. It is then elementary to see that the entire exponent becomes $-1 + \alpha(0) - 2p_2 \cdot p_4$ (upon using $\alpha' = 1$). On expressing this as $-2p_2 \cdot p_4 - c_3$, one concludes $c_3 = (1 - \alpha(0))$. N=5 is obviously a case where all intercepts are the same; nevertheless, $c_3 = 1 - \alpha(0) \neq 0$, contradicting Frampton's Eq.(2.57).

The vanishing of c_3 is *crucial* for the correctness of the Operator formalism as described by Frampton, which is generally accepted to be what underlies String Theory. In this interpretation of the Operator formalism, oscillators with the same dimensionality as the momentum vectors are introduced and it requires all c_n , $n \geq 3$ to vanish. Nambu, in [15] has devoted considerable attention to this question. He correctly points out that in general both c_2 and c_3 do not vanish. If c_3 vanishes but not c_2 , he argues for oscillators in one dimension higher than those of the momenta. But it is not at all clear if any consistent oscillator formalism exists if both c_2 , c_3 are non-zero. However, when $\alpha(0) = 1$, which is also equivalent to the Bootstrap condition, both c_2 and c_3 vanish. For higher N-point functions, where indeed the intercepts can in principle be different, it is obviously true that $c_n = 0$; $n \geq 4$ when all intercepts are equal, even if $\alpha(0) \neq 1$. We shall see later on that the choice $\alpha(0) = 1$ has many other desirable features.

These points will be crucial for the Operator formalism and String theory formulations. We shall elaborate on these when we look at 6-point and higher-point functions in the next section. A generalization of Eq.(15.57) can also be found in Eq.(4) of [15].

15.4 The Higher Point Functions

15.4.1 The 6-Point Function

In addition to 5-point functions analogous to the Veneziano formula, 6-point and 7-point functions were also quickly constructed, leading eventually to the N -point function for arbitrary N . The progress was amazingly fast! Chan Hong-Mo [16] was the first to obtain the 6-point function, and Chan and Tsun to obtain the 7-point function [17].

Once again, the construction of the 6-point function proceeds by (i) identifying an ordering and labelling of the six momenta p_i ; $i = 1, \dots, 6$, (ii) cyclic and anticyclic permutations are considered to be equivalent (this is analogous to the $B(x, y) = B(y, x)$ symmetry of the Euler B-function), (iii) identifying the distinct channels, also called *Partitions* by Chan and Tsun, and *non-crossing hooks* by Koba and Nielsen [18], and finally, (iv) imposing the *Duality Constraints* which ensure poles in only compatible channels. In addition, the independent channel variables have to be identified and a measure that ensures cyclic symmetry has to be incorporated. It is important to get a clear understanding of each of these steps as essentially the same ingredients are invoked for the general N -point construction too.

Let us take a look at the 6-point function construction of Chan in some detail. As already explained, there are six channels of the type $2 \rightarrow 4$, and three channels of the type $3 \rightarrow 3$. More explicitly, these are (1,2)(3,4,5,6), (2,3)(4,5,6,1), etc. (6 of them of the first type), and, (1,2,3)(4,5,6), (2,3,4)(5,6,1), (3,4,5)(6,1,2) of the second type.

In terms of the polygon recipe, one considers a hexagon with edges identified with the six momenta. The total number of diagonals that can be drawn is nine and these correspond to the nine channels just enumerated. From any vertex, a maximum of three diagonals can be drawn, and they are obviously non-intersecting and therefore represent compatible channels for simultaneous poles. In terms of the same graphical picture it is easy to see, for example, that channels (12) and (23) cannot have simultaneous poles, nor can the channels (12) and (234).

Chan introduces the variables $(u_1, u_2, \dots, u_6)$ for the first type and (v_1, v_2, v_3) for the second. The Duality constraints take the form

$$u_1 = 1 - u_6 v_2 v_1 \quad v_1 = 1 - v_2 v_3 u_1 u_4 \quad (15.59)$$

and their cyclic permutations which result in 6 relations of the first type, and 3 of the second. Though there appear to be 9 duality constraints in all, only 6 of them turn out to be independent (see [16] for details). These 6 independent relations can be solved in terms of 3 independent variables. In the $N = 5$ case, 5 duality constraint relations were solved in terms of only 2 independent variables. At this point we only point out, without explanation, that the number of independent variables is always $N - 3$.

The three independent variables in this case signify three simultaneous variables that are *compatible*. There are clearly many choices and Chan chooses (u_1, u_6, v_3)

to be one such. The solution of the duality constraints in these variables turns out to be

$$\begin{aligned} u_2 &= \frac{1-u_1}{1-u_1v_3}; u_3 = \frac{(1-v_3)(1-u_1u_5v_3)}{(1-u_1v_3)(1-u_5v_3)}; u_4 = \frac{1-u_5}{1-u_5v_3} \\ u_6 &= 1-u_1u_5v_3; v_1 = \frac{1-u_1v_3}{1-u_1u_5v_3}; v_2 = \frac{1-u_5v_3}{1-u_1u_5v_3} \end{aligned} \quad (15.60)$$

The Jacobian for the cyclic interchange of independent variables from (u_1, u_5, v_3) to (u_2, u_6, v_1) (analogous to the change $(f_3, f_5) \rightarrow (f_4, f_1)$ in the case of Virasoro's treatment of the 5-point function) is claimed by Chan to be

$$J(u_2, u_6, v_1; u_1, u_5, v_3) = \frac{u_1^2}{v_1} \cdot \frac{v_3}{u_6^2} \quad (15.61)$$

Using all these Chan writes down his expression for the 6-point function:

$$B_6^{Chan} = \int du_1 du_5 dv_3 \frac{v_3}{u_6^2} \prod_{i=1}^6 u_i^{x_i} \prod_{j=1}^3 v_j^{y_j} \quad (15.62)$$

where

$$\begin{aligned} x_1 &= -1 - \alpha_{12} = -1 - a_2 - (p_1 + p_2)^2; x_2 = -1 - a_2 - (p_2 + p_3)^2 \dots \\ y_1 &= -2 - \alpha_{234} = -2 - a_3 - (p_2 + p_3 + p_4)^2 \dots \end{aligned} \quad (15.63)$$

On comparing with Nambu's notation, one sees that $x_1 = (1, 2)$, etc., but $y_1 = -1 + (2, 3, 4)$, etc. Using all this, Chan's 6-point function can be rewritten as

$$\begin{aligned} B_6^{Chan2} &= \int du_1 du_5 dv_3 \frac{v_3}{(1-u_1u_5v_3)^2} u_1^{(12)} u_5^{(56)} v_3^{-1+(456)} \left(\frac{1-u_1}{1-u_1v_3} \right)^{(23)} \\ &\cdot \left[\frac{(1-v_3)(1-u_1u_5v_3)}{(1-u_1v_3)(1-u_5v_3)} \right]^{(34)} \left(\frac{1-u_5}{1-u_5v_3} \right)^{(45)} \left(\frac{1-u_1v_3}{1-u_1u_5v_3} \right)^{-1+(234)} \\ &\cdot \left(\frac{1-u_5v_3}{1-u_1u_5v_3} \right)^{-1+(345)} (1-u_1u_5v_3)^{(16)} \end{aligned} \quad (15.64)$$

We now compare this with what would obtain on using Eq. (15.54):

$$\begin{aligned} B_6^{Nam} &= \int \frac{dxdydz}{(1-xy)(1-yz)} x^{(12)} y^{(123)} z^{(1234)=(56)} \left(\frac{1-x}{1-xy} \right)^{(23)} \\ &\cdot \left[\frac{(1-y)(1-xyz)}{(1-xy)(1-yz)} \right]^{(34)} \left(\frac{1-z}{1-yz} \right)^{(45)} \\ &\cdot \left(\frac{1-xy}{1-xyz} \right)^{(234)} \left(\frac{1-yz}{1-xyz} \right)^{(345)} (1-xyz)^{(2345)=(16)} \end{aligned} \quad (15.65)$$

There appear to be two differences; the Jacobians (invariant measures), and the fact that the three-body terms like (123) are occurring with an additional -1 . Remarkably, the two differences compensate each other exactly to produce complete agreement between the two expressions after we make the identifications $(u_1 \rightarrow x)$, $(u_5 \rightarrow z)$, $(v_3 \rightarrow y)$. This can be seen on using

$$\frac{v_3}{u_6^2} \cdot v_1^{y_1} v_2^{y_2} v_3^{y_3} = \frac{1}{(1 - u_1 v_3)(1 - u_5 v_3)} v_1^{y_1+1} v_2^{y_2+1} v_3^{y_3+1} \quad (15.66)$$

We end this discussion by how the 6-point functions can be brought to the Bardakci-Ruegg form (see Eq. (2.55) of [4] for the 5-point function and his Eq. (2.65) for the N -point function, as well as Eq. (4) of [15]), by using simple tricks as done before. Since the two expressions, i.e. Chan's and Nambu's, we shall show it for Nambu's expression. It suffices to display just the integrand for Nambu's expression:

$$x^{(12)} y^{(123)} z^{(56)} (1-x)^{-2p_2 \cdot p_3 - c_2} (1-y)^{-2p_3 \cdot p_4 - c_2} (1-z)^{-2p_4 \cdot p_5 - c_2} \\ \cdot (1-xy)^{-2p_2 \cdot p_4 - c_3} (1-yz)^{-2p_3 \cdot p_5 - c_3} (1-xyz)^{-2p_2 \cdot p_5 - c_4} \quad (15.67)$$

15.4.2 N-point Function

Now we come to the construction of the N -point analog of the Veneziano amplitude. This followed the 5-point and 6-point constructions immediately. Chan and Tsun were the first to give a systematic construction [19], followed by Goebel and Sakita [20], Bardakci and Ruegg [21], and, Koba and Nielsen [18]. Hopkinson and Plahte [22] also established many recursion relations among the N -point functions. The most detailed and systematic treatments are to be found in the works of Chan and Tsun, and, Koba and Nielsen. It is important to stress that Koba and Nielsen explicitly assume the Bootstrap condition. We shall essentially follow Chan and Tsun because of the pedagogical ease of their approach. Frampton in his book also gives a clear discussion which is essentially the paper of Chan and Tsun.

Chan and Tsun identify the channels through the notion of *partitions* P which they carefully define as a “separation” of the N momenta into two groups *without* changing their order, with each group consisting of at least two momenta. It is easy to see that this is the same as the notion of “hooks” introduced by Koba and Nielsen, and the diagonals in the graphical method. From the polygon method, it is also easy to see that the total number of channels is $\frac{N(N-3)}{2}$. They introduce the notation $(i, j) \subset P$ for the partition $(i, i+1, \dots, j)(j+1, j+2, \dots, N, 1, 2, \dots, i-1)$, $i \geq 2, j > i$. Further, they introduce a precise notion of dual partitions $\bar{P} \supset (\bar{i}, \bar{j})$: two partitions $(i, j) \subset P$ and $(\bar{i}, \bar{j}) \subset \bar{P}$ are defined to be *dual* to each other if the sets of numbers $(i, \dots, j)$ and $(\bar{i}, \dots, \bar{j})$ intersect but do not contain each other. As an example, consider the partition (1, 2) for $N=4$; the partition (2, 3) intersects it with the intersection 2, but does not contain the former. These partitions are dual to each other.

In terms of these, Chan and Tsun find a simple and elegant generalization of the duality constraints, used for example, by Virasoro: *dual partitions cannot have simultaneous poles*. In terms of the “conjugate variables” u_P for partition P , this means if a channel corresponding to the partition P has a pole, i.e. if $u_P = 0$, then none of the partitions dual to P can have poles, i.e. $u_{\bar{P}} \neq 0$. They then write down the entire set of duality constraints in the compact form:

$$u_P = 1 - \prod_{\bar{P}} u_{\bar{P}} \quad (15.68)$$

More explicitly,

$$u_{i,j} = 1 - \prod_{p,q} u_{p,q} \prod_{r,s} u_{r,s} \quad 1 \leq p < i; i \leq q < j; i < r \leq j; j < s \leq N-1 \quad (15.69)$$

From the graphical method, it is also evident that there are $N-3$ independent channels which can have simultaneous poles (from any vertex a maximum of $N-3$ diagonals can be drawn). Chan and Tsun introduce the independent variables $u_{1,j}$ ($j = 2, \dots, N-2$) corresponding to them. Without loss of generality they define $u_{i,i} = 0$. The number of duality constraints is $\frac{N(N-1)(N-2)(N-3)}{24}$, a staggeringly large number as N increases. Nevertheless, Chan and Tsun, as well as Koba and Nielsen managed to solve them all in terms of the $N-3$ independent variables. The reason behind this apparent magic will become clear when we discuss the *Koba-Nielsen variables* and an underlying *Projective Invariance*. We follow the treatment of Chan and Tsun but encourage the reader to Koba and Nielsen’s extensive work.

The solution to the duality constraints found by Chan and Tsun is

$$u_{p,q} = \frac{(1 - u_{1,p} u_{1,p+1} \dots u_{1,q-1})(1 - u_{1,p-1} u_{1,p} \dots u_{1,q})}{(1 - u_{1,p-1} u_{1,p} \dots u_{1,q-1})(1 - u_{1,p} u_{1,p+1} \dots u_{1,q})} \quad (15.70)$$

with the restrictions

$$p = 2, \dots, N-2; q = 3, \dots, N-1; p < q \quad (15.71)$$

An issue immediately arises which Chan and Tsun have not addressed explicitly. From the allowed range of q values, the variable $u_{1,N-1}$ appears in the solution of the duality constraints (but $u_{1,N}$ never makes an appearance). Though this is not an independent variable, it looks like its actual value needs to be specified. We claim that it should be set equal to zero. A proof of this will naturally follow from the Koba-Nielsen variables to be discussed shortly.

The next ingredient is the invariant measure, equivalently the Jacobian for cyclic change of independent variables. We just quote Chan and Tsu’s results (their Eq. (9)):

$$J_1 = \prod_{i < j} (u_{i,j})^{j-i-1} \quad (15.72)$$

Frampton explicitly evaluates this in terms of the independent variables to be (see his Eq. (2.61)):

$$J_1 = \prod_{i=2}^{N-3} (1 - u_{1,i} u_{1,i+1}) \quad (15.73)$$

In this the condition $u_{1,N-1} = 0$ has to be used. This Jacobian is indeed of the form found by Nambu. Putting everything together, Chan and Tsun arrive at their N -point function:

$$B_N^{ChanTsun} = \int_0^1 \frac{1}{J_1} \prod_{j=2}^{N-2} du_{1,j} \prod_P (u_P)^{x_P} \quad (15.74)$$

where $x_{i,j} = -1 - \alpha_{i,j}$ with $\alpha_{i,j} = a_{j-i+1} + (p_i + p_{i+1} + \dots p_j)^2$. It should be noted that the channel factors that Chan and Tsun use now are uniformly of the type $x = -1 - \alpha(s)$ for *all* channels, whereas in their treatment of the 6-point function they had used $x = -1 - \alpha_2(s)$ for the two-body channels and $y = -2 - \alpha_3(s)$ for 3-body channels. We have already commented on how their Jacobian for the 6-point function was different from others, and how the two differences had exactly compensated each other.

Now we show how this expression for B_N can be brought into the form first introduced by Bardakci and Ruegg [21]. In particular how the exponent of $(1 - u_{1,i} \dots u_{1,j-1})$ simplifies when $j \geq i + 2$. On using Eq. (15.70), it is easy to see that this factor occurs exactly *once* in each of the four factors in the integrand of B_N :

$$u_{i,j}^{-1-\alpha_{i,j}} u_{i+1,j-1}^{-1-\alpha_{i+1,j-1}} u_{i,j-1}^{-1-\alpha_{i,j-1}} u_{i+1,j}^{-1-\alpha_{i+1,j}} \quad (15.75)$$

The said factor occurs in the numerator of the first two, and in the denominator of the last two. Consequently, the exponent of the factor is

$$-1 - \alpha_{i,j} - 1 - \alpha_{i+1,j-1} + 1 + \alpha_{i,j-1} + 1 + \alpha_{i+1,j} \quad (15.76)$$

and it is straightforward to simplify this to

$$-c_{j-i+1} - 2p_i \cdot p_j \quad (15.77)$$

resulting in the Bardakci-Ruegg form of the N -point function [21]:

$$B_N^{BaRu} = \int_0^1 \prod_{j=2}^{N-2} du_{1,j} u_{1,j}^{-1-\alpha_{1,j}} \prod_{2 \leq i < j \leq N-1} (1 - u_{1,i} \dots u_{1,j-1})^{-c_{j-i+1} - 2p_i \cdot p_j} \quad (15.78)$$

This is the same as Eq. (2.65) of [4]. Frampton goes on to claim in his Eq. (2.66) that when all the intercepts are equal, this reduces to:

$$B_N^{BaRuspl} = \int_0^1 \prod_{j=2}^{N-2} du_{1,j} u_{1,j}^{-\alpha_{1,j}-1} (1 - u_{1,j})^{\alpha(0)-1} \cdot \prod_{2 \leq i < j \leq N-1} (1 - u_{1,i} \dots u_{1,j-1})^{-2p_i \cdot p_j} \quad (15.79)$$

We have two criticisms of this claim: (i) the exponent of $(1 - u_{1,j})$ should be $(1 - \alpha(0))$, and, (ii) the $j = i + 2$ terms above must have a $-c_3$ in their exponent. Both are serious mistakes. However, both difficulties disappear for the case of $\alpha(0) = 1$. For the moment, we ignore the second difficulty related to c_3 , and state the correct form:

$$B_N^{BaRuspl} = \int_0^1 \prod_{j=2}^{N-2} du_{1,j} u_{1,j}^{-\alpha_{1,j}-1} (1 - u_{1,j})^{1-\alpha(0)} \cdot \prod_{2 \leq i < j \leq N-1} (1 - u_{1,i} \dots u_{1,j-1})^{-2p_i \cdot p_j} \quad (15.80)$$

15.4.3 Koba-Nielsen Variables

As we already remarked, it was somewhat miraculous that a very large number of duality constraints ($\frac{N(N-1)(N-2)(N-3)}{24}$) could be solved in terms of just $N - 3$ independent variables. Another issue, though not obviously related to this, is whether a manifestly cyclic representation can be found for B_N (so far, the cyclic symmetry though present is not manifest; one has to show that it emerges by virtue of the invariant measure chosen). Koba and Nielsen offered a brilliant resolution to both the questions in terms of what are now called *Koba-Nielsen Variables* and an associated *Projective Invariance* [23, 24]. We refer the reader to these original papers for all the details and the beautiful mathematics of projective invariance. Here we shall give an overview of the results. We also refer the reader to Frampton's book [4] for extensive discussions.

The idea is to place N points z_i , in cyclic order ($z_{N+1} = z_1$; $z_0 = z_N$), along an arbitrary circle in the complex plane (it is very important that the cyclic order be maintained while integrating over the z_i). Then the brilliant demonstration by Koba and Nielsen was that identifying the channel variables $u_{i,j}$ with the anharmonic ratios $(z_i, z_{i-1}; z_j, z_{j+1})$ where the anharmonic ratio $(a, b; c, d)$ is defined by

$$(a, b; c, d) \equiv \frac{(a-c)(b-d)}{(a-d)(b-c)} \rightarrow u_{i,j} = (z_i, z_{i-1}; z_j, z_{j+1}) \quad (15.81)$$

and that such an identification solves *all* the duality constraints at once! The proof that this is so can also be found in Eqns.(2.85)-(2.90) of Frampton's book.

Before proceeding further, it should be noted that this identification automatically satisfies $u_{i,i} = 0$ which Chan and Tsun had taken by definition. We had also pointed out the need to set $u_{1,N-1} = 0$. This too follows from the Koba-Nielsen prescription:

$$u_{1,N-1} = (z_1, z_0 = z_N; z_{N-1}, z_N) = \frac{(z_1 - z_{N-1})(z_N - z_N)}{(z_1 - z_N)(z_N - z_{N-1})} = 0 \quad (15.82)$$

The next step is to express the N -point function itself in terms of the Koba-Nielsen variables. The Bardakci-Ruegg form is particularly suited for this purpose. Before attempting that, we explicitly work out the 4-point function to bring out many important subtleties. We start with one of our earlier representations of the 4-point function:

$$B_4 = \int_0^1 du_{1,2} u_{1,2}^{-\alpha_{1,2}-1} u_{2,3}^{-\alpha_{2,3}-1} \quad (15.83)$$

where, as before, $\alpha_{1,2} = a_2 - 2 + 2 p_1 \cdot p_2$, and, $\alpha_{2,3} = a_2 - 2 + 2 p_2 \cdot p_3$ (we have made use of $\alpha' \mu^2 = -1$). The Koba-Nielsen identification for $u_{1,2}$ and $u_{2,3}$ are

$$\begin{aligned} u_{1,2} &= (z_1, z_4; z_2, z_3) = \frac{(z_1 - z_2)(z_3 - z_4)}{(z_1 - z_3)(z_2 - z_4)} \\ u_{2,3} &= (z_2, z_1; z_3, z_4) = \frac{(z_2 - z_3)(z_1 - z_4)}{(z_2 - z_4)(z_1 - z_3)} \end{aligned} \quad (15.84)$$

We have rewritten the expressions so that all differences $z_i - z_j$ are such that $j > i$, for reasons that will become clear soon. On using both the explicit expressions as well as directly by the identities satisfied by anharmonic ratios (see Eq. (2.82) [4]), it follows that $u_{1,2} + u_{2,3} = 1$ so that the above representation for B_4 coincides with the earlier ones.

The independent variable $u_{1,2}$ depends on four variables, z_1, z_2, z_3, z_4 . How can this be if we are to eventually look for a representation where the Koba-Nielsen variables are themselves to be used as integration variables? Also, what should be the correct measure to be used? Koba-Nielsen gave an extensive analysis of this important issue based on the concept of the Haar measure in Group Theory. The key to both these is the invariance of the anharmonic ratios under the so-called *Möbius* transformation:

$$z \rightarrow z' = \frac{az + b}{cz + d} \quad ad - bc = 1 \quad (15.85)$$

The most general Möbius transformation is a composition of (i) *scaling* transformation $z' = cz$, (ii) followed by *translation* $z'' = z' + d = cz + d$, followed by (iii) *inversion* $z''' = \frac{-1}{cz+d}$ and by a suitable final translation $z'''' = z''' + \frac{a}{c}$. It is easy to show that these transformations form a 3-parameter Non-abelian group. This feature will play a major role in discussions to come. The invariance of the anharmonic ratio under the Möbius transformations follows from the fact that being made up of

only differences of z 's it is invariant under translations, being homogeneous of zero degree means it is invariant under scaling, and finally the invariance under inversion is guaranteed by the fact that in the anharmonic ratio each z_i occurs exactly once in the numerator and once in the denominator. For the last aspect, it is enough if each z_i occurs the same number of times in the numerator as in the denominator. This way more complicated projective invariants can be constructed.

Because of this invariance, it is possible to fix any three of the four z_i leaving, in this case, only one free variable. It is usual to fix z_1, z_{N-1}, z_N ; therefore we shall treat z_1, z_3, z_4 as fixed and only treat z_2 as the integration variable. It is then elementary to show

$$du_{1,2} = -dz_2 \frac{(z_1 - z_4)(z_3 - z_4)}{(z_1 - z_3)(z_2 - z_4)^2} \quad (15.86)$$

This is rewritten in the following way, for reasons to be explained shortly.

$$\frac{\prod_{i=1}^4 dz_i}{dz_1 dz_3 dz_4} (z_1 - z_3)(z_3 - z_4)(z_4 - z_1) \prod_{k=1}^4 \frac{1}{(z_k - z_{k+2})} \quad (15.87)$$

As $u_{1,2}$ is a projective-invariant variable, the above is a *Projective Invariant Measure*!

From a mathematical point of view, division by infinitesimals as suggested above is nonsensical. But, as Koba and Nielsen have stressed, in the present context, three of the four variables $z_i, i = 1, \dots, 4$ can be fixed because of projective invariance and the volume element can be written as the product of the volume element for the subspace that is fixed times the volume element for the rest (in the case of 4-point function, this is one dimensional). Projective invariance also implies that the 3-volume can be pulled outside the integral which in effect is realized by the division by infinitesimals. We now turn to expressing the integrand I_4 in terms of the KN-variables:

$$I_4 = \left[\frac{(z_1 - z_2)(z_3 - z_4)}{(z_1 - z_3)(z_2 - z_4)} \right]^{-2p_1 \cdot p_2 - c_2} \cdot \left[\frac{(z_2 - z_3)(z_1 - z_4)}{(z_2 - z_4)(z_1 - z_3)} \right]^{-2p_2 \cdot p_3 - c_2} \quad (15.88)$$

where $c_2 = a_2 - 1$ (as stressed before, this disagrees with Frampto's c_2 in sign, and the correct sign is crucial from several points). Now due to the overall energy-momentum conservation $p_1 + p_2 + p_3 + p_4 = 0$, one gets the dependencies $p_1 \cdot p_2 = p_3 \cdot p_4$, and, $p_1 \cdot p_4 = p_2 \cdot p_3$ (there are more!). Using them, the numerator of the above can be cast in the form:

$$N = (z_1 - z_2)^{-2p_1 \cdot p_2 - c_2} (z_2 - z_3)^{-2p_2 \cdot p_3 - c_2} \cdot (z_3 - z_4)^{-2p_3 \cdot p_4 - c_2} (z_1 - z_4)^{-2p_4 \cdot p_1 - c_2} \quad (15.89)$$

Each term is of the *nearest neighbour* type. The denominator, on using the less obvious relation $2p_1 \cdot p_2 + 2p_2 \cdot p_3 = -2\mu^2 - 2p_1 \cdot p_3$ along with the obvious $p_1 \cdot p_3 = p_2 \cdot p_4$, can be rewritten (on using $\mu^2 = -1$) as

$$D^{-1} = \prod_{k=1}^4 (z_k - z_{k+2}) (z_1 - z_3)^{2c_2 - 2p_1 \cdot p_3} (z_2 - z_4)^{2c_2 - 2p_2 \cdot p_4} \quad (15.90)$$

Putting together all the terms, the integrand can be expressed as

$$I_4 = \prod_{k=1}^4 (z_k - z_{k+2}) \prod_{1 \leq i < j \leq 4} (z_i - z_j)^{-2p_i \cdot p_j - c_j - i + 1} \quad (15.91)$$

with the identifications $c_2 = a_2 - 1$; $c_3 = -2c_2$; $c_4 = c_2$. These c 's are what are given by the earlier formula $c_n = a_n - 2a_{n-1} + a_{n-2}$ on noting the *periodicities* of a_n , i.e. $a_i = a_{N-i}$ for the N -point case, which translates to $a_3 = a_1 = 1$; $a_4 = a_0 = 1$. As the periodicity conditions explicitly involve N , it would have been less confusing to label all a_i and consequently all c_n by an additional label of N .

Before analysing this further, we rewrite the integrand displaying separately the terms that depend explicitly on the momenta, on c_2 , and the rest:

$$I_4 = \prod_{m=1}^4 (z_m - z_{m+2}) \left[\prod_{i=1}^4 \frac{(z_i - z_{i+1})}{(z_i - z_{i+2})} \right]^{1-a_2} \prod_{1 \leq i < j \leq 4} (z_i - z_j)^{-2p_i \cdot p_j} \quad (15.92)$$

Strictly speaking, the first product of terms should actually be considered as being part of the momentum-dependent terms. That also follows from the observation that the integrand is in itself a projective invariant being made up entirely of $u_{i,j}$, and the factor depending on c_2 is also projective invariant from the arguments given earlier.

We have taken the trouble of giving a detailed step-by-step derivation of this as we have disagreements with various claims in the literature; these issues being very important and the reader should have a clear appreciation of them. First of these is with regard to c_2 itself; Frampton, at several places in his book, has claimed this to be $1 - a_2$ as against our $a_2 - 1$. A related issue is what happens when $c_2 \neq 0$. Both Frampton (see Eq. (2.100) of [4]) and even more explicitly Fairlie [25] (see the quoted expression for the N -point amplitude which he ascribes to [23]) say that only nearest neighbour terms are induced by $a_2 \neq 1$ (or $c_2 \neq 0$). But we clearly see even for the simplest 4-point function that $c_2 \neq 0$ leads to not only nearest neighbour terms of the type $(z_i - z_{i+1})$ but also next-nearest neighbour terms like $(z_i - z_{i+2})$. That this has to be so followed directly from Projective Invariance, without the need for any explicit calculations. The c_2 -dependent terms have to be invariant under projective transformations on their own, and only nearest-neighbour terms cannot be projective invariant by themselves. Our explicit calculation of Eq. (15.92) is also in conformity with this. We shall give similar explicit results for the 5-point function also.

Now the first product of terms of I_4 exactly cancels an identical product of terms in $du_{1,2}$ to give the final expression for B_4 in terms of the KN-variables:

$$B_4 = \int \frac{\prod_{i=1}^4 dz_i}{dV_{134}} \prod_{k=1}^4 \theta(z_k - z_{k+1}) \left[\prod_{i=1}^4 \frac{(z_i - z_{i+1})}{(z_i - z_{i+2})} \right]^{1-a_2} \prod_{1 \leq i < j \leq N} (z_i - z_j)^{-2p_i \cdot p_j} \quad (15.93)$$

where

$$dV_{abc} \equiv \frac{dz_a dz_b dz_c}{(z_a - z_b)(z_b - z_c)(z_c - z_a)} \quad (15.94)$$

and a, b, c are any three of the N -points. Explicit factors of θ -functions have been introduced to underscore the important requirement that the cyclic order of z_i should be maintained (see also Fairlie and Jones [26]). The meaning of the θ -function is the corresponding θ -function for angles in case z_i are arranged on a circle. When the circle is conformally mapped to the real line, the θ -function has the usual meaning. This is precisely the result of [23, 24] (see Eq. (5.9) of the second reference), not what is mentioned in [4, 25].

It is important to notice that the relevant range has now become $i : 1, \dots, N - 1; j = 2, \dots, N$, in place of the earlier $i : 2, \dots, N - 2; j : 3, \dots, N - 1$. As in all earlier cases, the full scattering amplitude $A_4(s, t)$ is to be obtained from B_4 by adding the contributions from *all* the permutations inequivalent to cyclical ones. Therefore, $A_4(s, t) = B_4(s, t) + B_4(t, u) + B_4(u, s)$, while the cyclic invariance of the measure only guaranteed $B_4(x, y) = B_4(y, x)$.

The course of the discussion at this point dramatically depends on whether $c_2 = 0$, equivalently $a_2 = 1$, or not. When $a_2 \neq 1$, the integrand explicitly depends on cyclic ordering. Even when $a_2 = 1$, it may still seem to depend on the cyclic ordering because of the $\theta(z_k - z_{k+1})$ terms. But when the sum over all inequivalent permutations are done to arrive at $A_4(s, t)$, when $a_2 = 1$, the effect is as if only a single expression for B_4 is considered, but with the new proviso that the integration variable z_2 can take *unrestricted* values on the circle (see also [4]). More specifically, when $a_2 = 1$, all factors in the integrand except the θ -functions can be trivially summed with appropriate change of variables, whose net effect is to remove all the θ -functions. Then a single B_4 will actually be invariant under *all* permutations of the invariants like $p_i \cdot p_j$. The reader is referred to the work of Fairlie and Jones [26] for further details. As emphasized before, $a_2 = 1$ or $c_2 = 0$ is what one gets if one assumes the Bootstrap condition.

We simply quote the results for the 5-point function without working through all the details. Now there are all together 5 KN-variables, and projective invariance is used to fix z_1, z_4, z_5 leaving z_2, z_3 as the free variables, whose values are however restricted by the cyclic ordering. The independent channel variables are $u_{1,2}$ and $u_{1,3}$ which take on the values of the anharmonic ratios $(z_1, z_5; z_2, z_3)$ and $(z_1, z_5; z_3, z_4)$ respectively. The integration measure was $\frac{du_{1,2} du_{1,3}}{u_{2,4}}$. The KN-expression for $u_{2,4}$ is $(z_2, z_1; z_4, z_5)$. To express the measure in terms of $dz_2 dz_3$ one has to compute the Jacobian $\frac{\partial(u_{1,2}, u_{1,3})}{\partial(z_2, z_3)}$:

$$du_{1,2} du_{1,3} = \frac{\partial(u_{1,2}, u_{1,3})}{\partial(z_2, z_3)} dz_2 dz_3 \quad (15.95)$$

This is greatly facilitated by the fact that $u_{1,3}$ does not depend on z_2 . Hence the Jacobian becomes

$$\frac{\partial(u_{1,2}, u_{1,3})}{\partial(z_2, z_3)} = \frac{\partial u_{1,2}}{\partial z_2} \cdot \frac{\partial u_{1,3}}{\partial z_3} \quad (15.96)$$

so it's essentially the same type of computation for the $N = 4$ case but done twice (instead of a 2×2 determinant!).

In fact, this feature persists as we move to higher N -point functions where the channel variables are $u_{1,j}$, $j = 2, \dots, N-2$ whose KN-variable description (they are $z_1, z_2, \dots, z_N$ in cyclic order with z_1, z_{N-1}, z_N fixed by projective invariance) are $(z_1, z_N; z_j, z_{j+1})$. Therefore, $u_{1,j}$ does not depend on the independent KN-variables $z_i : i = 2, \dots, j-1$. Consequently, the Jacobian matrix is *banded upper triangular* (with exactly two bands, to be precise). It's determinant is simply the product of it's diagonal elements:

$$\frac{\partial(u_{1,2}, \dots, u_{1,N-2})}{\partial(z_2, \dots, z_{N-2})} = \prod_{j=2}^{N-2} \frac{\partial u_{1,j}}{\partial z_j} \quad (15.97)$$

The expression for the projective invariant and cyclic invariant measure is

$$\frac{du_{1,2} du_{1,3}}{u_{2,4}} = \prod_{i=1}^5 \frac{dz_i}{z_i - z_{i+2}} dV_{145}^{-1} \quad (15.98)$$

This is the same form we obtained in the $N = 4$ case also and is in full agreement with the general result given by Frampton in [4] (see his Eq. (2.97)). The integrand I_5 is also easily worked out:

$$I_5 = \prod_{k=1}^5 (z_k - z_{k+2}) \prod_{1 \leq i < j \leq 5} (z_i - z_j)^{-2p_i \cdot p_j - c_{j-i+1}} \quad (15.99)$$

This is also of the same structure as the one we got for the $N = 4$ case. Putting everything together, along with the θ -functions to enforce the cyclic orderings, one gets

$$B_5 = \int \frac{\prod_{k=1}^5 dz_i}{dV_{abc}} \prod_{k=1}^5 \theta(z_k - z_{k+1}) \prod_{1 \leq i < j \leq 5} (z_i - z_j)^{-2p_i \cdot p_j - c_{j-i+1}} \quad (15.100)$$

Exactly identical in structure to the $N = 4$ case and in conformity with the general N case given in Eq. (2.98) of [4] (it does not explicitly include the theta-functions):

$$B_N = \int \frac{\prod_{k=1}^N dz_i}{dV_{abc}} \prod_{k=1}^N \theta(z_k - z_{k+1}) \prod_{1 \leq i < j \leq N} (z_i - z_j)^{-2p_i \cdot p_j - c_{j-i+1}} \quad (15.101)$$

Beyond $N = 5$, more than one type of intercept becomes necessary and it would be an interesting exercise to work out the generic case, but we shall not do so here. When all the intercepts are equal,

$$B_N = \int \frac{\prod_{k=1}^N dz_i}{dV_{abc}} \prod_{k=1}^N \theta(z_k - z_{k+1}) \left[\prod_{l=1}^N \frac{(z_l - z_{l+1})}{(z_l - z_{l+2})} \right]^{1-a_2} \prod_{1 \leq i < j \leq N} (z_i - z_j)^{-2p_i \cdot p_j} \quad (15.102)$$

This is also the form obtained by Koba and Nielsen explicitly. The full amplitude A_N is obtained by summing B_N over all inequivalent permutations of the particle labels. As in the cases of $N = 4, 5$, as long as $a_2 \neq 1$ there is no simple way of doing this. But when $a_2 = 1$, the summation over inequivalent permutations effectively removes the restrictions to cyclic orderings, equivalently by removing the theta-functions to give, for the full A_N a single integral of remarkable simplicity:

$$A_N = \int \frac{\prod_{k=1}^N dz_i}{dV_{abc}} \prod_{1 \leq i < j \leq N} (z_i - z_j)^{-2p_i \cdot p_j} \quad (15.103)$$

This single integral is invariant under *all* permutations of the invariants. Frampton (see his remarks after his Eq. (3.55) [4]) states this in the more powerful language of symmetries: when $a_2 = 1$, the integrand becomes fully symmetric under *all* the permutations ($N!$ of them) of the pairs (z_i, p_i) . In other words, there is a significant *symmetry enhancement*! Of course, upon integrating the z_i , the full permutation symmetry survives among the invariants like $p_i \cdot p_j$.

References

1. G. Veneziano, *Nuovo Cimento* **57A**, 190 (1968)
2. J.E. Mandula, R.S. Slansky, *Phys. Rev. Lett.* **20**, 1402 (1968)
3. H. Fujisaki, *Prog. Theor. Phys.* **43**(1), 101 (1970)
4. P.H. Frampton, *Dual Resonance Models* (Benjamin, W.A., 1974)
5. M. Ademollo, H.R. Rubinstein, G. Veneziano, M.A. Virasoro, *Phys. Rev. Lett.* **19**, 1402 (1967)
6. M. Ademollo, H.R. Rubinstein, G. Veneziano, M.A. Virasoro, *Phys. Lett.* **27B**, 99 (1968)
7. S. Mandelstam, *Phys. Rev.* **166**, 1539 (1968)
8. M. Ademollo, H.R. Rubinstein, G. Veneziano, M.A. Virasoro, *Phys. Rev.* **176**, 1904 (1968)
9. H.R. Rubinstein, A. Schwimmer, G. Veneziano, M.A. Virasoro, *Phys. Rev. Lett.* **21**, 491 (1968)
10. K. Huang, Formalism and phenomenology of complex angular momentum, in *Summer School in Elementary Particle Physics* (Brookhaven National Lab., 1969) (BNL 50212), July 22–Aug 29, 1969
11. L.M.B.C. Compos, F.S.R.P. Cunha, J. Inequal. Special Funct. **3**, 1–28 (2012)
12. K. Bardakci, R. Ruegg, *Phys. Lett.* **28B**, 342 (1968)
13. M.A. Virasoro, *Phys. Rev. Lett.* **22**, 37 (1968)
14. K. Bardakci, R. Ruegg, *Phys. Lett.* **28B**, 671 (1969)
15. Y. Nambu, Quark Model and the factorization of the Veneziano amplitude, reprinted in *Broken Symmetry: Selected Papers of Y. Nambu*, ed. by T. Eguchi, K. Nishijima (World Scientific, 1995)
16. Chan Hong-Mo, *Phys. Lett.* **28B**, 425 (1969)
17. C. Hong-Mo, T. Tsun, CERN Preprint TH969 (1968)
18. Z. Koba, H.B. Nielsen, *Nuc. Phys.* **B10**, 633 (1969)

19. Chan Hong-Mo, T. Tsun, Phys. Lett. **28B**, 485 (1969)
20. C.J. Goebel, B. Sakita, Phys. Rev. Lett. **22**, 257 (1969)
21. K. Bardakci, R. Ruegg, Phys. Rev. **181**, 1884 (1969)
22. J.F.L. Hopkinson, E. Plahte, Phys. Lett. **28B**, 489 (1969)
23. Z. Koba, H.B. Nielsen, Nuc. Phys. **B12**, 517 (1969)
24. Z. Koba, H.B. Nielsen, Z. F. Phys. **229**, 243 (1969)
25. D.B. Fairlie, The birth of string theory, in *String Theory and Fundamental Interactions*. Lecture Notes on Physics, vol. 737
26. D.B. Fairlie, K. Jones, Nuc. Phys. **B15**, 323 (1970)

The Operator Formalism and The Dual Resonance Model

16

16.1 Introduction

In this chapter we discuss the *Operator Formalism* for Dual Resonance Models. Recall our discussion in Chap. 15 of *precise duality* which emerged as a property of the Veneziano amplitude. This was given a mathematically well-defined meaning as expressed in Eq. (15.21). It is important to notice that this statement in itself was given entirely in terms of (narrow) resonances. Such models are called *Dual Resonance Models*. Clearly, the N -point generalizations of the Veneziano model, also discussed extensively in Chap. 15, are all cases of such models.

As it turns out there is an underlying quantum (harmonic) oscillator structure to the N -point functions. We shall not go into a detailed historical and logical discussion of this issue, but only mention some highlights. The earliest works that had a bearing on this issue were those of Fubini, Veneziano and Gordon [1,2] during their investigations of the level structure of Dual Resonance Models. Independently, Susskind was pioneering the study of the inter-relationship between Veneziano model (amplitudes) and harmonic oscillators [3–5]. Nambu used the oscillator structure to show the very important property of *Factorizability* of the N -point amplitudes [6]. For a good discussion of the oscillator framework in dual resonance models, the reader is referred to the review article by Paolo Di Vecchia [7]. For most part, we shall follow the treatment in Frampton's book [8]. It is to be emphasized that Di Vecchia and Frampton use opposite metric conventions.

Nambu first considers the general case when c_n (see Chap. 15 for details) are non-vanishing. He begins by examining the simplest case of the 4-point function with non-vanishing c_2 , i.e. when $a_2 \neq 1$. Nambu found that as long as $c_2 \neq 0$, a *five-dimensional* oscillator structure was needed (with the metric $(- - - + -)$) to reproduce the Veneziano amplitude. As one goes to higher N -point functions, particularly for $N \geq 6$, we saw that many different types of Regge intercepts arise and that even in the special limit of all the intercepts becoming equal, the various

c_n 's can still differ in signs. In such cases, it is not at all clear that a consistent five-dimensional oscillator formulation is possible (see Nambu's remarks after his Eq. (14) of [6]). However, when all the intercepts are equal, it turns out that four-dimensional oscillators suffice, as discussed in detail in [8]. We shall only discuss this. Many important issues like *ghosts*, the degeneracy of states, etc. are tied up with the precise oscillator structure.

Before proceeding, we recall the expression for the N-point function in the Bardakci-Ruegg form of Eq. (15.80):

$$B_N = \int_0^1 \prod_{j=2}^{N-2} du_{1,j} u_{1,j}^{-\alpha_{1,j}-1} (1-u_{1,j})^{1-\alpha(0)} \prod_{2 \leq i < j \leq N-1} (1 - \prod_{k=i}^{j-1} u_{1,k})^{-2p_i \cdot p_j} \quad (16.1)$$

This is valid only when all the intercepts are equal, without them necessarily being equal to unity. We shall address the very special features of the unit intercept as we go along.

16.2 Operator Formalism-I

16.2.1 Preliminaries

In order to appreciate the essence of the operator formalism for the dual resonance models, consider the case of a single quantum harmonic oscillator described by the algebra of creation and destruction operators:

$$[a, a^\dagger] = 1 \quad (16.2)$$

Along with a vacuum state satisfying $a|0\rangle = 0$. Then on using the famous *Baker-Campbell-Hausdorff* identity

$$e^A \cdot e^B = e^B e^A e^{[A,B]} \quad (16.3)$$

which is valid when $[A, [A, B]] = [B, [B, A]] = 0$, as indeed happens in the harmonic oscillator case considered, it is easily shown that

$$\langle 0 | e^{\beta^* a} \cdot e^{\alpha a^\dagger} | 0 \rangle = e^{\beta^* \alpha} \quad (16.4)$$

for arbitrary complex (α, β) . The BCH formula is well known and is textbook material now. It's most general form, i.e. when $[A, [A, B]]$ and $[B, [B, A]]$ do not vanish (see Eq. (C.16) of [9]) is

$$e^A \cdot e^B = e^{A+B + \frac{[A,B]}{2} + \frac{[A,[A,B]] + [B,[B,A]]}{12} + \dots} \quad (16.5)$$

This is a very important relation in studies of Group Theory (see [9] Appendix C). The introduction of the oscillator *Coherent States* defined by

$$|\alpha\rangle = e^{\alpha a^\dagger} |0\rangle \quad (16.6)$$

with arbitrary complex α will greatly facilitate the rest of the discussion. The reader should note that we have not *normalized* the coherent states as the normalization factor does not play any role in our considerations. It immediately follows from Eq. (16.4) that

$$\langle \beta | \alpha \rangle = e^{\beta^* \alpha} \quad (16.7)$$

We shall refer to this in future as the *overlap* condition. An obvious consequence, to be referred to as the *parametric shift* condition, is

$$e^{\beta a^\dagger} |\alpha\rangle = |\alpha + \beta\rangle \quad (16.8)$$

modulo some normalization factors which we are ignoring for reasons mentioned above. On using the easily provable and well-known operator identity (see Eq. (8.29) of [10])

$$e^A B e^{-A} = B + [A, B] + \frac{[A, [A, B]]}{2} \dots \quad (16.9)$$

it is easily seen that coherent states are *eigenstates* of the annihilation operator:

$$a |\alpha\rangle = \alpha |\alpha\rangle \quad (16.10)$$

We shall refer to this as the *eigenvalue* condition. Here too the normalization factors are unimportant. A non-trivial property that will be very important later on is

$$\beta a^\dagger a |\alpha\rangle = |\beta \alpha\rangle \quad (16.11)$$

which we shall refer to as *evolution*. We will leave it to the reader to prove this.

So the basic idea is to realize the $-2p_i \cdot p_j$ terms in the various exponents of the N -point function as arising out of the inner-products of appropriate coherent functions or equivalently, as arising out of appropriate oscillator algebras. It is easy to see that the oscillators have to be d -dimensional if the momenta p_i are d -dimensional, and that they must also carry the Minkowski signature. We shall basically follow the treatment in Sect. 2.6 of [8] except that he restricts his analysis to $d = 4$. The required oscillator algebra is

$$[a_\mu^{(n)}, a_\nu^{(m)\dagger}] = -g_{\mu\nu} \delta_{mn} \quad g_{\mu\mu} = (+, -, -, \dots) \quad (16.12)$$

all other components of the metric being zero, and no Einstein summation convention is implied. We shall make one more modification; Frampton treats the annihilation and creation operators as *contravariant* four vectors. As it will turn out later in the

context of a string interpretation, these operators turn out to be the mode expansions of a field that plays the role of a *coordinate*. Accordingly, we take them to be *covariant* and write

$$[a_{(n)}^\mu, a_{(m)}^\nu]^\dagger = -g^{\mu\nu} \delta_{mn} \quad g^{\mu\mu} = (+, -, -, \dots) \quad (16.13)$$

In addition to these, it becomes necessary to introduce the so called *zero-mode* oscillators satisfying the standard Heisenberg algebra:

$$[\hat{Q}^\mu, \hat{P}^\nu] = -i g^{\mu\nu} \quad (16.14)$$

The coherency parameters $(\alpha, \beta \dots)$ will accordingly be *contravariant* vectors $(\alpha_{\mu,n}, \beta_{\mu,n} \dots)$. Though coherent states can be associated with each individual oscillator of the type

$$|\alpha_{\mu}^{(n)}\rangle = e^{\alpha_{\mu}^{(n)} a_{(n)}^\mu} |0\rangle \quad (16.15)$$

where no summation is carried out over either μ or n , due to Lorentz covariance it is more useful to introduce coherent states of the type

$$|\{\alpha\}\rangle = e^{\sum_n \alpha_{\mu,n} a_{(n)}^\mu} |0\rangle \quad (16.16)$$

where the μ -indices are also summed over. For these classes of coherent states, the generalizations of the above-discussed properties are:

$$\langle \{\beta\} | \{\alpha\} \rangle = e^{-\sum_{n,\mu\nu} \beta_{\mu}^{(n)*} \alpha_{\nu}^{(n)} g^{\mu\nu}} = e^{-\beta^{(n)*} \cdot \alpha^{(n)}} \quad (16.17)$$

The negative sign in comparison to the single oscillator case should be carefully noted.

Likewise, the eigenvalue statement generalizes to

$$a_{(n)}^\nu |\{\alpha\}\rangle = -\alpha_{(n)}^\nu |\{\alpha\}\rangle \quad \alpha_{(n)}^\nu = g^{\mu\nu} \alpha_{\mu,n} \quad (16.18)$$

Lastly

$$x^{-\sum_n n a_{(n)}^\dagger \cdot a_{(n)}} |\{\alpha\}\rangle = |\{\alpha_x\}\rangle \quad \alpha_{x\mu}^{(n)} = x^n \alpha_{\mu}^{(n)} \quad (16.19)$$

16.2.2 Oscillators and N-Point Functions

First, one defines a *Vertex Operator* (see [8] Eq. (2.117)):

$$V(p) = e^{i\sqrt{2} p \cdot \sum_{n=1}^{\infty} \frac{a_{(n)}^\dagger}{\sqrt{n}}} e^{i\sqrt{2} p \cdot \sum_{n=1}^{\infty} \frac{a_{(n)}}{\sqrt{n}}} \quad (16.20)$$

and a *Propagator* (see [8] Eq. (2.118))

$$D(s) = \int_0^1 dx x^{-\alpha(s)-1+R} (1-x)^{1-\alpha(0)} \quad R = -\sum_{n=1}^{\infty} n a_{(n)}^{\dagger} \cdot a_{(n)} \quad (16.21)$$

Again, the error in the sign of $\alpha(0)$ in [8] is to be carefully noted. In terms of these constructs, the central result of the operator formalism for dual models can be stated as

$$B_N = \langle 0 | V(p_2) D(s_{12}) V(p_3) D(s_{123}) \dots V(p_{N-1}) | 0 \rangle \quad (16.22)$$

where B_N is the N -point function as stated in Eq. (16.1). $|0\rangle$ is the vacuum state annihilated by *all* the annihilation operators. As a check, Frampton evaluates the 4-pt function

$$\begin{aligned} B_4 &= \langle 0 | V(p_2) D(s) V(p_3) | 0 \rangle \\ &= \int_0^1 dx x^{-\alpha(s)-1} (1-x)^{1-\alpha(0)} \langle 0 | V(p_2) x^R V(p_3) | 0 \rangle \end{aligned} \quad (16.23)$$

In this and all subsequent evaluations, Frampton advocates commuting all x^R -like factors to one side first. This, though correct, is tedious and error-prone. We shall follow a different method which proves to be easy and straightforward even for arbitrary N , and which makes optimal use of the properties of coherent states listed above. Towards this, note

$$V(p)|0\rangle = |i\sqrt{2}p\{\frac{1}{\sqrt{n}}\}\rangle \quad (16.24)$$

in an obvious notation. Once again, the normalization factors for the coherent states have been ignored. With this, the 4-point function becomes

$$B_4 = \int_0^1 dx x^{-\alpha(s)-1} (1-x)^{1-\alpha(0)} \langle -i\sqrt{2}p_2\{\frac{1}{\sqrt{n}}\} | x^{-R} | i\sqrt{2}p_3\{\frac{1}{\sqrt{n}}\} \rangle \quad (16.25)$$

On using all the properties of coherent states above,

$$\begin{aligned} &\langle -i\sqrt{2}p_2\{\frac{1}{\sqrt{n}}\} | x^{-\sum_n n a_{(n)}^{\dagger} \cdot a_{(n)}} | i\sqrt{2}p_3\{\frac{1}{\sqrt{n}}\} \rangle \\ &= \langle -i\sqrt{2}p_2\{\frac{x^n}{\sqrt{n}}\} | i\sqrt{2}p_3\{\frac{1}{\sqrt{n}}\} \rangle = e^{2p_2 \cdot p_3 \sum_n \frac{x^n}{n}} \\ &= (1-x)^{-2p_2 \cdot p_3} \end{aligned} \quad (16.26)$$

On further noting $\alpha(t) = \alpha(0) - 2 + 2p_2 \cdot p_3$, one recovers for B_4 the standard Veneziano form. Crucial to this is the correct sign of $\alpha(0)$ in our Eq. (16.1)!

We now show how the 6-pt function can also be evaluated with equal ease. Let us start with the operator identity for this case

$$B_6 = \langle 0|V(p_2) D(s_{12}) V(p_3) D(s_{123}) V(p_4) D(s_{1234}) V(p_5)|0\rangle \quad (16.27)$$

Writing this out in full and making use of our Eq. (16.24),

$$B_6 = \int \prod_{k=1}^3 dx_k (1-x_k)^{1-\alpha(0)} x_1^{-\alpha(s_{12})-1} x_2^{-\alpha(s_{123})-1} x_3^{-\alpha(s_{56})-1} X \quad (16.28)$$

where

$$\begin{aligned} X &= \langle -i\sqrt{2}p_2\{\frac{1}{\sqrt{n}}\}|x_1^{-R} V(p_3) x_2^{-R} V(p_4) x_3^{-R}|i\sqrt{2}p_5\{\frac{1}{\sqrt{n}}\}\rangle \\ &= e^{2p_4 \cdot p_5 \sum_n \frac{x_3^n}{n}} \cdot e^{2p_2 \cdot p_3 \sum_n \frac{x_1^n}{n}} \\ &\quad \cdot \langle -i\sqrt{2}p_2\{\frac{x_1^n}{\sqrt{n}}\}|e^{i\sqrt{2}p_3 \cdot \sum_n \frac{a(n)}{\sqrt{n}}} x_2^{-R} e^{i\sqrt{2}p_4 \cdot \sum_n \frac{a^\dagger(n)}{\sqrt{n}}} |i\sqrt{2}p_5\{\frac{x_3^n}{\sqrt{n}}\}\rangle \\ &= (1-x_3)^{-2p_4 \cdot p_5} (1-x_1)^{-2p_2 \cdot p_3} \\ &\quad \cdot \langle -i\sqrt{2}p_2\{\frac{x_1^n}{\sqrt{n}}\} - i\sqrt{2}p_3\{\frac{x_2^n}{\sqrt{n}}\}|x_2^{-R}|i\sqrt{2}p_5\{\frac{x_3^n}{\sqrt{n}}\} + i\sqrt{2}p_4\{\frac{x_2^n}{\sqrt{n}}\}\rangle \\ &= (1-x_3)^{-2p_4 \cdot p_5} (1-x_1)^{-2p_2 \cdot p_3} \\ &\quad \cdot \langle -i\sqrt{2}p_2\{\frac{x_1^n}{\sqrt{n}}\} - i\sqrt{2}p_3\{\frac{x_2^n}{\sqrt{n}}\}|i\sqrt{2}p_5\{\frac{x_2^n x_3^n}{\sqrt{n}}\} + i\sqrt{2}p_4\{\frac{x_2^n}{\sqrt{n}}\}\rangle \\ &= (1-x_1)^{-2p_2 \cdot p_3} (1-x_3)^{-2p_4 \cdot p_5} (1-x_2)^{-2p_3 \cdot p_4} \\ &\quad \cdot (1-x_1 x_2)^{-2p_2 \cdot p_4} (1-x_2 x_3)^{-2p_3 \cdot p_5} (1-x_1 x_2 x_3)^{-2p_2 \cdot p_5} \end{aligned} \quad (16.29)$$

It is easily seen that this correctly reproduces the earlier 6-pt function.

We have shown every step explicitly to help the reader see how to do this for arbitrary N-point functions. To repeat, the sequence of steps is, i. convert the initial and final vacuum states to coherent states, ii. apply *evolution* equations to both of these, iii. use *eigenvalue* conditions, iv. use *parametric shifts*, v. use *evolution* and finally, vi. the *overlap* condition. This will be the same sequence for arbitrary N. We shall leave it as an exercise for the reader to verify Eq. (16.22).

One important point has been overlooked and that is the fact that B_N must necessarily be accompanied by a delta-function expressing overall energy-momentum conservation. This is accomplished by introducing additional oscillators of the standard quantum-mechanical type $[q_\mu, p_\nu] = i g_{\mu\nu}$, also called the *zero-mode* oscillators, and by enlarging the vacuum state from $|0\rangle$ to $|0, 0\rangle$ where the second $|0\rangle$ is acted upon by only the zero-mode oscillators. We refer the reader to [7, 11] for details, but provide a summary here. Enlarging the oscillators this way was first done by Fubini and Veneziano in [12]. These zero-mode oscillators play a larger role than in just accounting for overall energy-momentum conservation.

They started by constructing a *field* out of the oscillators in much the same way as was discussed in Chap. 6. A key difference now is that the modes are labelled by integers (n) to be contrasted with their continuous labelling by $\mathbf{k}$ in the free-field case considered there. It turns out to be more natural to introduce fields labelled by the Koba-Nielsen complex variable z . Fubini and Veneziano introduced the operators (named after them):

$$Q^\mu(z) = Q^{\mu,+}(z) + Q^{\mu,0}(z) + Q^{\mu,-}(z) \quad (16.30)$$

with

$$Q^{(+)}(z) = i\sqrt{2} \sum_{n=1}^{\infty} \frac{a_{(n)}}{\sqrt{n}} z^{-n}; \quad Q^{(-)}(z) = -i\sqrt{2} \sum_{n=1}^{\infty} \frac{a_{(n)}^\dagger}{\sqrt{n}} z^n; \quad Q^{(0)}(z) = \hat{Q} - 2i \ln z \hat{P} \quad (16.31)$$

The zero-mode operators satisfy the standard Heisenberg algebra, $[\hat{Q}_\mu, \hat{P}_\nu] = i g_{\mu\nu}$. Their introduction necessitates extending the ground state to $|0, 0\rangle$ with the second zero referring to the ground state of the zero-mode system. In particular, the momentum eigenstates

$$|0, P\rangle = e^{-P \cdot \hat{Q}} |0, 0\rangle \quad \langle 0, P' | 0, P\rangle = (2\pi)^D \delta^{(D)}(P - P') \quad (16.32)$$

will play an important role.

Fubini and Veneziano generalize the earlier introduced vertex operators to

$$V(z; p) \equiv e^{ip \cdot Q^{(-)}(z)} \cdot e^{ip \cdot \hat{Q}} \cdot e^{2p \cdot \hat{P} \ln z} \cdot e^{ip \cdot Q^{(+)}(z)} \quad (16.33)$$

The relationship of these vertex operators to what was introduced earlier is

$$V(1; p) = e^{ip \cdot \hat{Q}} \cdot V(p) \quad (16.34)$$

We have set $\alpha' = -1$ as in our earlier treatments. A remarkable property of the new vertex functions (see Eq. (64) of [7]) is

$$\langle 0, 0 | \prod_{j=1}^N V(z_i; p_i) | 0, 0 \rangle = (2\pi)^D \delta^{(D)} \left(\sum_{i=1}^N p_i \right) \prod_{i>j} (z_i - z_j)^{-2p_i \cdot p_j} \quad (16.35)$$

Consequently,

$$(2\pi)^D \delta^{(D)} \left(\sum_{i=1}^N p_i \right) B_N = \int \frac{\prod_{i=1}^N dz_i \theta(z_i - z_{i+1}) (z_i - z_{i+1})^{\alpha(0)-1}}{dV_{abc}} \cdot \langle 0, 0 | \prod_{i=1}^N V(z_i; p_i) | 0, 0 \rangle \quad (16.36)$$

where dV_{abc} is the projective-invariant measure introduced by Koba and Nielsen. Some clarifications are in order at this stage. All the $e^{ip \cdot \hat{Q}}$ factors from each vertex function can be gathered to one end, and then with the help of Eq. (16.32), one gets the overall energy-momentum conservation delta function. The second remark is that the N -point function can now be obtained entirely in terms of the new vertex operators without the need for any *propagators*! The zero-mode part $e^{-2p \cdot \hat{P} \ln z}$ part of the new vertex operators is essential for this.

16.2.3 Factorizability and Degeneracies

Nambu [6] introduced the oscillator formalism to display the very important property of *Factorizability* of the dual amplitudes as expressed by the N -point functions. Factorizability is the property of amplitudes whereby the residues at the Regge poles are factorizable. He also introduced the vertex operators of Eq. (16.20) though he did not introduce the propagators of Eq. (16.21). Another important difference of Nambu's approach lay in the introduction of *five-dimensional* oscillators in contrast to what has been used so far in this chapter. He used the 5-dimensional oscillators even to discuss the 4-point function. We shall restrict ourselves to the use of only 4-dimensional oscillators. Frampton, however, stresses the importance of the 5-dimensional oscillators and in particular their relevance for the so-called *Anomaly* term in his Eqs. (3.163–3.193), but we shall not get into those technicalities.

We follow the treatment given by Frampton to demonstrate factorizability simplified by the assumption of $\alpha(0) = 1$. In that sense, the proof of factorizability is less general than the one given by Nambu but is sufficient to bring out clearly how factorizability works. Let us rewrite the operator expression for the N -point function as

$$B_N = \langle 0 | V(p_2) D(s_{1,2}) V(p_3) D(s_{1,3}) \dots V(p_j) D(s_{1,j}) V(p_{j+1}) \dots V(p_{N-1}) | 0 \rangle \quad (16.37)$$

where we have introduced a somewhat more systematic notation for the invariants $s_{i,j} = (p_i + p_{i+1} + \dots + p_j)^2$. Now let us consider the pole of $\alpha_{1,j}(s_{1,j})$ at $\alpha_{1,j} = M$ with M an integer. When $\alpha(0) = 1$, the integral in the propagator $D(s_{1,j})$ can be explicitly carried out:

$$D(s_{1,j}) = \int_0^1 du_{1,j} u_{1,j}^{-\alpha_{1,j}(s_{1,j})-1+R} = \frac{1}{R - \alpha_{1,j}(s_{1,j})} \quad (16.38)$$

Thus the poles occur at the eigenvalues of R . The corresponding eigenstates are generically degenerate. In terms of these degenerate eigenstates, $R|M, i\rangle = M|M, i\rangle$ with the additional label i keeping track of the degeneracies. Letting $d(N)$ denote the degree of degeneracy, one can introduce the orthonormal basis $\langle M, i | M, j \rangle = \delta_{ij}$ and $\sum_{j=1}^{d(M)} |M, j\rangle \langle M, j| = 1$. The basis is of course not unique.

Inserting the completeness relation (for specific M) twice in B_N and using the orthogonality relations,

$$B_N = \sum_{i=1}^{d(M)} \langle 0 | V(p_2) D(s_{1,2}) V(p_3) D(s_{1,3}) \dots V(p_j) | M, i \rangle \frac{1}{M - \alpha_{1,j}(s_{1,j})} \cdot \langle M, i | V(p_{j+1}) \dots V(p_{N-1}) | 0 \rangle \quad (16.39)$$

demonstrating factorizability. More precisely, the residue at the pole is expressible as a *sum* of factors. This treatment is easily extended to multiple poles as they may arise in appropriate N-point functions.

Though we have displayed the states $|M, i\rangle$ as the eigenstates of R with eigenvalue M , the pole itself is at $\alpha_{1,j}(s_{1,j}) = M$. This suggests introducing the operator $L_0 = R - \alpha_{1,j}(s_{1,j})$ whose eigenvalue at the poles is exactly 1. There are many significances to L_0 , as will be elaborated as we go along.

We end this discussion with some remarks about $d(M)$. We shall treat the simpler case of only one dimension, as generalization to arbitrary space-time dimensions is straightforward as each dimension acts independently of all others. In that case, $R = \sum_r r a_r^\dagger a_r$ (we have changed the notation to minimize confusion!). Denoting the eigenvalues of $a_r^\dagger a_r$ by m_r , the degeneracy $d(M)$ is simply the number of ways $\sum_r r m_r$ can take the value M . In this instance, the eigenvalues of $a_r^\dagger a_r$ are integers and non-degenerate. It is easy to see that

$$\prod_{r=1}^{\infty} (1 - x^r)^{-1} = \prod_r \sum_{m_r=0}^{\infty} x^{r m_r} = \sum_{M=0}^{\infty} d(M) x^M \quad (16.40)$$

When the space-time dimensions are D , the extension of this *generating function* for $d(M)$ is obvious:

$$\prod_{r=1}^{\infty} (1 - x^r)^{-D} = \sum_{M=0}^{\infty} d(M) x^M \quad (16.41)$$

The asymptotic behaviour of $d(M)$ for large M was solved in the celebrated work of Hardy and Ramanujan on *Partitions of Integers* in number theory [13]. The leading term of their result is

$$d(M) \simeq_{M \rightarrow \infty} c(D) M^{-\frac{D+3}{4}} e^{2\pi \sqrt{\frac{DM}{6}}} \quad (16.42)$$

This analysis is purely combinatorial, but in the context of the dual models $d(M)$ can be the degeneracy of states only if the important assumption of *linear independence* is made. This means that there are no linear dependences among the eigenstates with given M . Such dependences would naturally *reduce* the degeneracy. As we shall see later, such dependencies are crucial to eliminate various unphysical features like *ghosts*, etc.

It is important to also remark here that the above considerations did not invoke any specific models, nor even general reasonings based on statistical mechanics. Hagedorn [14], even before the dual resonance models, had considered exponentially increasing degeneracies. Nambu, during discussion session of his talk [6] also alludes to statistical mechanical basis for such growth, but does not refer to Hagedorn.

16.3 Operator Formalism-II

In this section we shall address some more technical aspects of the operator formalism for dual resonance models. Due to the vastness of the topic, we shall only highlight some major aspects by virtue of their importance both for the development of conventional string theories, as well as for the effective string theories. Both Frampton in [8] as well as Di Vecchia [7] have extensively covered all relevant aspects.

We begin by looking at the operator formalism for the Koba-Nielsen projective invariance we discussed in Chap. 15. This was a 3-parameter continuous (Lie) group acting on the Koba-Nielsen variables z_i via the Möbius transformation:

$$z \rightarrow z' = \frac{az + b}{cz + d} \quad ad - bc = 1 \quad (16.43)$$

They form the group $SU(1,1)$. The three generators L_{-1}, L_0, L_1 satisfy the Lie Algebra:

$$[L_1, L_{-1}] = 2L_0 \quad [L_0, L_{\pm 1}] = \mp L_{\pm 1} \quad (16.44)$$

$SU(1,1)$ is a *non-compact* group in contrast to the compact group $SU(2)$. The Lie algebra of the latter is

$$[J_+, J_-] = 2J_3 \quad [J_3, J_{\pm}] = \pm J_{\pm} \quad (16.45)$$

The $SU(1,1)$ Lie algebra admits both finite-dimensional matrix representations as well as infinite-dimensional representations of the form:

$$L_0 = -z \frac{d}{dz} \quad L_{-1} = -\frac{d}{dz} \quad L_1 = -z^2 \frac{d}{dz} \quad (16.46)$$

For the operator formalism of the dual models, one has to find the realization of the $SU(1,1)$ generators on the Fock-space generated by the oscillator creation and annihilation operators. These turn out to be [8]:

$$L_0 = -\sum_{n=1}^{\infty} n a_{(n)}^{\dagger} \cdot a_{(n)} - \hat{P}^2; \quad L_1 = -\sum_{n=1}^{\infty} \sqrt{n(n+1)} a_{(n)}^{\dagger} \cdot a_{(n+1)} + i\sqrt{2} \hat{P} \cdot a_{(1)} \quad (16.47)$$

with $L_{-1} = L_1^{\dagger}$. That these operators satisfy the Lie algebra of Eq. (16.44) can be verified on using Eq. (16.13). The projective invariance of the N -point functions when expressed in terms of the Koba-Nielsen variables z_i was true for any value

of the common intercept $\alpha(0)$. We also discussed how the projective invariance got promoted to a much larger invariance when $\alpha(0) = 1$. Henceforth, we shall only be discussing the $\alpha(0) = 1$ case.

This enhanced invariance motivated Fubini and Veneziano [15] to introduce the generators for an infinite-dimensional *Generalized Projective Group*:

$$L_n = -z^{n+1} \frac{d}{dz} \quad (16.48)$$

as obvious generalizations of Eq. (16.46). The Lie algebra of these generators is easily worked out:

$$[L_m, L_n] = (m - n) L_{m+n} \quad (16.49)$$

also called a *Virasoro Algebra*. These authors, however, did not construct these generators in terms of the oscillators as done for the SU(1,1) case (see Eq. (16.47)). That was done by Clavelli and Weis [16], who, however, did not publish their central results! The oscillator representation they proposed was

$$\begin{aligned} L_m = & - \sum_{n=1}^{\infty} \sqrt{n(n+m)} a_{(n)}^{\dagger} \cdot a_{(n+m)} + i\sqrt{2m} \hat{P} \cdot a_{(m)} \\ & + \frac{1}{2} \sum_{n=1}^{m-1} \sqrt{n(m-n)} a_{(n)} \cdot a_{(m-n)} \end{aligned} \quad (16.50)$$

with $L_{-m} = L_m^{\dagger}$. Here, $\hat{P}$ is the total momentum operator. Again, the algebra of these generators can be worked out easily, and they had many surprises:

$$[L_m, L_n] = (m - n) L_{m+n} + \frac{D}{12} m(m^2 - 1) \delta_{m,-n} \quad (16.51)$$

Among the several surprises, (i) the algebra is not strictly a Lie algebra because of the extra term, which however vanishes for the cases $m = 0, \pm 1$. The latter is of course the SU(1,1) subalgebra. The extra term is variously called *Central Extension*, *Anomaly Term*, etc. It is a purely quantum effect. So the algebra is actually a *Centrally Extended* Lie algebra, as the additional term, being a c-number, commutes with all the generators L_m , (ii) for the first time the space-time dimensionality D appears explicitly! This will have radical ramifications, as will be seen soon.

Frampton points out (see Eqs. (3.66)–(3.71) of [8]) that the structure of the anomaly term follows from very general algebraic principles, without the need for explicit operator evaluations. We reproduce his arguments here: (a) the Jacobi identity requires

$$[[L_n, L_m], L_p] + [[L_p, L_n], L_m] + [[L_m, L_p], L_n] = 0 \quad (16.52)$$

Writing the centrally extended algebra as

$$[L_m, L_n] = (m - n) L_{m+n} + C_{m,n} \quad (16.53)$$

it is seen that the above Jacobi identity implies for $C_{m,n} = -C_{n,m}$:

$$(n-m)C_{n+m,p} + (m-p)C_{m+p,n} + (p-n)C_{p+n,m} = 0 \quad (16.54)$$

(b) The fact that the $SU(1,1)$ subalgebra receives no central extension implies

$$C_{1,-1} = C_{1,0} = C_{-1,0} = 0 \quad (16.55)$$

all other C 's vanishing because of the antisymmetry $C_{n,m} = -C_{m,n}$. Now the trick is to choose (n, m, p) in such a way that exactly one of the three terms vanishes in Eq. (16.54). There are many ways of doing this but not all of them are useful. For example, the choice $m = 0$, $p = -n$ leads to $n C_{n,-n} + n C_{-n,n} = 0$ which holds anyway because of antisymmetry. Frampton uses $m = -1$, $p = 1 - n$ which, on using $C_{1,-1} = 0$ yields

$$\frac{C_{n,-n}}{C_{n-1,-n+1}} = \frac{n+1}{n-2} \quad (16.56)$$

This can be solved iteratively to give

$$C_{n,-n} = \frac{n(n^2-1)}{6} C_{2,-2} \quad (16.57)$$

While this determines the “diagonal” entries like $C_{n,-n}$, it has not proved that “off-diagonal” entries like $C_{n,m}$, for $n \neq -m$ vanish. We leave that as an exercise to the reader. This general algebraic proof does not fix $C_{2,-2}$ to be $\frac{D}{2}$. That must be fixed by explicit operator calculation of $[L_2, L_{-2}]$. However, in that computation only the c-number term needs to be calculated. It is worth mentioning that makes sense only when a particular ordering of the creation and annihilation operators in L_0 is specified. This is usually taken to be the so-called *Normal Ordering* whereby all creation operators are to the left of all the annihilation operators.

16.4 Physical States of the Dual Resonance Models

In much of the discussions about the Veneziano formula and its generalizations to N -point functions, the only states that figured were the particle states in the initial and final stages of scattering, and by the Bootstrap condition, these were required to lie on the Regge trajectories that entered the description of the scattering. For spin-less case, we saw that this demanded the particles to be tachyonic, though some would take issues with such particles as being *Physical*. Though tachyonic, these states were certainly of *positive* norm with positive energies, features essential for physical states. But with the demonstration of factorization above, the degenerate states $|M, i\rangle$ entered the picture, and one of the most important issues was whether all of them were physical. The very same questions arose in the subsequent developments of String Theories too. Because of their extreme importance, we will carefully elucidate the issues and the techniques used to resolve them.

With the set of oscillators $a_{(n)}^\mu, a_{(n)}^{\mu\dagger}$, a rather large number of *Harmonic Oscillator* states of the type

$$|\lambda_1, \lambda_2, \dots\rangle = \prod_{n=1}^{\infty} \frac{(a_{(n)}^{\mu_n})^{\lambda_{n;\mu_n}}}{\sqrt{\lambda_{n;\mu_n}}} |0\rangle \quad (16.58)$$

can be constructed. Because of the Minkowski signature, states created by *odd* numbers of time-like oscillators are *Negative Norm* states, also called *Ghosts*. These are clearly unacceptable physically. To see this, consider a single-mode vectorial oscillator with the algebra

$$[a^\mu, a^{\nu\dagger}] = -g^{\mu\nu} \quad g^{\mu\mu} = (-, +, +, \dots) \quad (16.59)$$

Clearly the state $a^{0\dagger}|0\rangle$ has *negative norm*, while those excited by the spatial components, $a^{i\dagger}|0\rangle$, have positive norm. The same situation arises, for example, in Quantum Electrodynamics, when the quantum vector-field $A_\mu(x)$ is mode-expanded as

$$A_\mu(x) = \int \frac{d^3k}{(2\pi)^3 2k_0} [a_\mu(k)e^{-ikx} + a_\mu^\dagger(k)e^{ikx}] \quad (16.60)$$

The algebra of the creation and annihilation operators of QED is the same as in Eq. (16.59), but for each mode labelled by $\mathbf{k}$. Therefore in QED, superficially there is one negative norm state (ghost) for every mode. A big challenge faced by the quantization of the electromagnetic field in this *Covariant* approach was the elimination of all such ghosts and show that the *Physical* state space only consists of *positive-norm* states as required of a consistent theory. The key to this resolution in QED is of course the recognition that $A_\mu(x)$ are not physically observable (even in the classical theory) by virtue of their *redundancy* due to the invariance of the theory under the so-called gauge transformations $A_\mu(x) \rightarrow A_\mu(x) + \partial_\mu \lambda(x)$ for *arbitrary* $\lambda(x)$. In the classical theory, this redundancy, if need be, can be removed by fixing the gauge, for example, by requiring $\partial_\mu A^\mu(x) = 0$. But in quantum theory, it is impossible to impose the operator constraint $\partial_\mu A^\mu(x) = 0$ on physical states as the positive-frequency part of this “constraint”, made with the annihilation operators, does not commute with the corresponding negative frequency part made with the creation operators. It is customary to impose the positive-frequency part $\partial_\mu A^{\mu,+} = 0$ on physical states. It then follows that this procedure suffices to remove all ghosts from the theory. This will be the guiding spirit for the dual models also.

It is worth making a few observations on the fock space of QED created by the creation operators. For clarity, let us consider only the one-particle states but generalization to arbitrary multiparticle states is possible, though somewhat tedious. The states, for each mode labelled by $\mathbf{k}$, are four of which one is negative norm and three are the positive norm. Of the three positive norm states, the *longitudinal* mode also decouples leaving behind exactly two physical states characteristic of the two transverse polarization states of the electromagnetic field (this is in $D = 4$; in arbitrary D , the number of transverse states is $D - 2$). Considerations such as these

will soon be seen to play a crucial role in deciding the physical states of the dual model also.

Before proceeding further, let us revisit the factorizability proof of Eq. (16.39) which was based on the assumption of $\alpha(0) = 1$ leading to the very simple form for the propagator $D(s) = \frac{1}{R-\alpha(s)}$. Let us, for the moment, relax the $\alpha(0) = 1$ requirement. The propagator will no longer be of the simple form above. Nevertheless, Eq. (16.37) can be rewritten as

$$B_N = \langle p_{2,j} | D(s_{1,j}) | p_{j+1,N-1} \rangle \quad (16.61)$$

where

$$\begin{aligned} |p_{j+1,N-1}\rangle &= V(p_{j+1})D(s_{1,j+1})V(p_{j+2})D(s_{1,j+2}) \dots V(p_{N-1})|0\rangle \\ \langle p_{2,j}| &= \langle 0|V(p_2)D(s_{1,2})V(p_3)D(s_{1,3}) \dots V(p_j) \end{aligned} \quad (16.62)$$

The projective invariance of B_N requires

$$W_1 \equiv (L_1 - L_0) |p_{j+1,N-1}\rangle = 0 \quad (16.63)$$

with an equivalent statement for $\langle p_{2,j}|$. We refer the reader to [7] (his Eqs. (92)–(95)) for a proof based on representations of states like $|p_{k,l}\rangle$ explicitly in terms of the Koba-Nielsen variables.

So, are the conditions of Eq. (16.63) the analogs of the gauge-constraints on physical states in QED, and do they suffice to remove all the ghosts from the spectrum of dual models? The answer is “No” to both! There are time-direction oscillators for every mode labelled by (n) in the dual model and one would require at least as many gauge-like constraints for each (n) . More importantly, the above conditions are more like invariances under “global” gauge transformations while what are required to remove all the ghosts are like the infinitely many “local” gauge transformations.

However when $\alpha(0) = 1$, the propagator takes the form $D = \frac{1}{L_0-1}$ and it turns out there are infinitely many “gauge” conditions of the type:

$$W_n \equiv (L_n - L_0 - n + 1) |p_{1,M}\rangle = 0 \quad (16.64)$$

These were first obtained by Virasoro in [17] where he proved such “subsidiary conditions” when $\alpha(0) = 1$. He also showed in that work that these additional conditions were enough to eliminate all time-component oscillators. A straightforward way to prove these conditions is to use the $[L_n, L_0] = nL_n$ along with $[L_n, V(p)]$ to show

$$W_n V(p) = V(p)(W_n + n) \quad (W_n + n)D = (L_0 + n - 1)^{-1} W_n \quad (16.65)$$

along with the ground state conditions

$$(L_0 - 1)|0\rangle = L_n|0\rangle = 0 \quad n > 0 \quad (16.66)$$

16.4.1 Varieties of States of the Dual Model

Let us recall the expression for the residue of the pole at $\alpha_{1,j}(s_{1,j}) = M$:

$$\sum_{i=1}^{d(M)} \langle p_{(2,j)} | M, i \rangle \cdot \langle M, i | p_{j+1, N-1} \rangle \quad (16.67)$$

The naive degeneracy $d(M)$ was already discussed. At this point the states $|M, i\rangle$ obviously include both positive and negative norm states, and, as we shall see shortly even *zero-norm* states. Di Vecchia includes the total momentum P in the labelling of the eigenstates, and we do the same from this point onwards $|M, i\rangle \rightarrow |M, i, P\rangle$. With Di Vecchia's metric conventions the on-shell condition is $1 - \alpha' P^2 = M$, and with his $R = \sum_n n a_{(n)}^\dagger \cdot a_{(n)}$, L_0 is still 1 on shell.

Spurious States: the first important concept is that of the so-called spurious states. These are states that do not contribute to the residue (see Eq. (16.67)). Such states occur even when $\alpha(0) \neq 1$ though they are far fewer in that case. Because of the projective invariance condition Eq. (16.63), it is clear that any state $|M, i, P\rangle$ of the form $W_1^\dagger |\mu, P\rangle$ will be spurious. Of course $|\mu, P\rangle$ cannot be totally arbitrary as the spurious state still has to be an eigenstate of R , equivalently, of L_0 with eigenvalue 1.

Coming to $\alpha(0) = 1$ case, all states of the type $|s, P\rangle = W_m^\dagger |\mu, P\rangle$ are spurious. Following Di Vecchia (his Eqs. (154), (155)) the states $|\mu, P\rangle$ are given by

$$R|\mu, P\rangle = m_\mu |\mu, P\rangle; L_0 |\mu, P\rangle = (1 - m) |\mu, P\rangle; 1 - \alpha' P^2 = M \quad (16.68)$$

with $m = M - m_\mu$ (there is an error in Di Vecchia's Eq. (155)). The on-shell condition $(L_0 - 1)|\mu, P\rangle = -m|\mu, P\rangle$ yields, on using $W_m = L_m - m - (L_0 - 1)$,

$$|s, P\rangle = L_m^\dagger |\mu, P\rangle \quad (16.69)$$

which is a particularly useful representation of spurious states. The next step is in *defining* the physical states of the theory. According to Di Vecchia, physical states are defined to be those that are *orthogonal* to *all* the spurious states:

$$\langle s, P | phys \rangle = \langle \mu, P | L_m | phys \rangle = 0 \rightarrow L_m | phys \rangle = 0 \quad m \geq 1 \quad (16.70)$$

Combining with on-shell condition, the physical states satisfy

$$L_m |M, i, P\rangle = 0 \quad (L_0 - 1) |M, i, P\rangle = 0 \quad (16.71)$$

Once again, a comparison with QED is in order. Here too it would have been impossible to demand that all the generators L_m , for both positive and negative m annihilate physical states as the commutator $[L_m, L_{-m}] = 2mL_0 + \frac{D}{12}m(m^2 - 1)$ is non-vanishing. These very important conditions were first found by Del Giudice and Di Vecchia [18] and also by Yoshimura [19].

It is possible for states to be simultaneously spurious as well as physical in the sense of Eq. (16.71)! Let us consider such a spurious state $|s, P\rangle$ also satisfying the physical state conditions $L_m |s, P\rangle = 0 = (L_0 - 1)|s, P\rangle$. It is immediately obvious that such states must have zero-norm:

$$\langle s, P | s, P \rangle = \langle s, P | L_m^\dagger | \mu, P \rangle = 0 \quad (16.72)$$

As spurious states, these zero-norm states do not contribute to the residue and are in that sense not physical. There are also states called *conjugate* for which the signs of the time-like oscillators are changed. Collectively denoting (M, i, P) by λ , the completeness relation at a given mass shell (eigenvalue of R) takes the form

$$1 = \sum_{\lambda_+} |\lambda_+\rangle \langle \lambda_+| - \sum_{\lambda_-} |\lambda_-\rangle \langle \lambda_-| + \sum_{\mu} [|\mu\rangle \langle \mu_c| + |\mu_c\rangle \langle \mu|] \quad (16.73)$$

where $\lambda_{\pm}$ are respectively the positive and negative norm physical states, $|\mu\rangle$ are the spurious-physical, hence zero norm states, and $|\mu_c\rangle$ the corresponding conjugate states. As already explained, the last two terms never contribute to any residue. Thus, what is left is to show that the $|\lambda_-\rangle$ states also do not contribute, and in that sense they too are spurious and physical, though their norms are not zero.

16.4.2 Absence of Ghosts and DDF Construction

Ghosts can manifest in two different ways; one, as states with negative norms and two, when the partial-wave expansions have negative coefficients. In an earlier chapter we already noted that the Veneziano amplitudes can lead to ghosts of the second kind when $D > 26$. Now we turn to the issue of proving the absence of ghosts of the first kind in the operator formalism of dual models. The proof is quite involved and technical. We will not go into all the details but only highlight the major conceptual aspects. The reader is referred to [7] for details.

It is based on the works of Del Giudice, Di Vecchia and Fubini [20] and is called the DDF-formalism. The starting point is the construction of the so-called DDF operators $A_{n;i}$, $i = 1, \dots, D-2$, $D-2$ operators for each mode. They are designed to satisfy the relations

$$[A_{n;i}, L_m] = 0 \quad [A_{n;i}, A_{m;j}] = n \delta_{ij} \delta_{n+m,0} \quad (16.74)$$

The second of these is like oscillators for spatial directions and hence excites only positive-norm states while the first ensures that the states created by them obey the physical state conditions.

The outcome of the DDF analysis is that dual models are ghost-free in $D < 26$. This was also the conclusion of the partial-wave analysis. 26 dimensions appear in a variety of contexts and are called the *Critical Dimension* $D_c = 26$. The Fock-space constructed out of DDF operators also reveals another aspect of D_c ; it is only in 26 dimensions that the DDF states form a complete set. Exactly in 26 dimensions

some states become zero-norm states which decouple. The absence of ghosts in dual models for $D < D_c = 26$ was proved independently by Browser [21] and by Goddard and Thorn [22].

16.4.3 QED Revisited

Frampton provides a nice analysis of the physical states of QED in parallel to the analysis above in terms of spurious states, etc. [8]. He looks at the one-photon states of a given momentum $\mathbf{k}$. Naively there are four states (in $D = 4$) excited by the creation operators of which the one excited by the time-component has negative norm, while the three excited by the spatial operators are positive norm. Yet the physical states are only two corresponding to the transverse degrees of freedom. The massless nature of photons is crucial.

The gauge-condition $\partial_\mu A^\mu(x) = 0$ in momentum space takes the form $q_\mu A^\mu(q) = 0$, and the positive-frequency part, analogous to L_m for positive m becomes $L = q^\mu a_\mu$. With the choice $(1, 0, 0, 1)$ for q , this is $L = a_0 - a_3$. The spurious state in question is $(a_0^\dagger - a_3^\dagger)|0\rangle$, which is clearly a zero-norm state. The conjugate state is then $(a_0^\dagger + a_3^\dagger)|0\rangle$, which is also a *Null* state. Both these decouple for essentially the same reasons elaborated for the dual model, leaving behind the two positive norm states $a_1^\dagger|0\rangle, a_2^\dagger|0\rangle$ as the two transverse physical states.

16.4.4 First Two Excited States

We now discuss two very nice examples analysed by [7]. Although all the essentials are mentioned by them, we shall fill in the details to make it easier for the reader to appreciate fully these important examples. Di Vecchia considers the first two excited states, equivalently the $M = 1$ and $M = 2$ excited states where $\alpha(s) = M$. It is easy to see that the $M = 1$ state has $P^2 = 0$ and $M = 2$ state has $P^2 = 1$ (remember the ground state with $M = 0$ is tachyonic $P^2 = -1$). These follow simply on using $\alpha(0) = 1$.

Let us consider the massless $M = 1$ states. They are excited by the $a_{(1)}^\mu{}^\dagger$ operators and the most general state is of the form

$$|\psi_1\rangle = \epsilon_\mu a_{(1)}^\mu{}^\dagger |0, P\rangle \quad (16.75)$$

The naive degeneracy, in the sense of not taking into account any linear dependences, is D , and this would be the $D(1)$ introduced earlier. This is just the number of parameters in $|\psi_1\rangle$. But the physical state conditions $L_m|\psi_1\rangle = 0$ do introduce linear dependences. It suffices to only satisfy the $L_1|\psi_1\rangle$ physical state condition, which yields $\epsilon \cdot P = 0$. So far, it is looking exactly like the gauge-invariant description of one-photon states with the polarization ϵ satisfying the gauge-invariance restriction.

Now what about the remaining physical state conditions $L_m|\psi\rangle = 0$ for positive m greater than one? Firstly note the relation

$$[L_2, \epsilon \cdot a_{(1)}^\dagger] = \epsilon \cdot a_{(1)} \quad (16.76)$$

Because of this it follows that $L_2|\psi_1\rangle = 0$ due to $L_2|0, P\rangle = 0$. The rest of the physical state conditions for $m > 2$ can be shown to hold on using $[L_{m-1}, L_1] = (m-2)L_m$ iteratively. The masslessness $P^2 = 0$ allows one to choose a frame in which $P_\mu = (P, 0, 0, \dots, P)$. In that frame, the invariant condition $\epsilon \cdot P = 0$ yields $\epsilon_0 = \epsilon_{D-1} = \epsilon$. The most general physical state then takes the form:

$$|phys_1\rangle = \epsilon_i a_{(1)}^{i\dagger} |0, P\rangle + \epsilon(a_{(1)}^{0\dagger} + a_{(1)}^{D-1\dagger}) |0, P\rangle \quad i = 1, \dots, D-2 \quad (16.77)$$

Superficially, the degeneracy has been reduced by 1 to $D-1$, but there is more! Also, the time-component oscillator has not been eliminated. Of the states above, the first term denotes $D-2$ transverse states of positive norm. But the second term is of zero-norm and can in fact be written as $L_1^\dagger |0, P\rangle$ and is therefore orthogonal to all the physical transverse modes. In other words, it is a spurious state that decouples, bringing down the true degeneracy to $D-2$.

This analysis also adds further meaning to the $\alpha(0) = 1$ condition which has so far been introduced purely algebraically without attempting to attach any physical significance to it. Quite obviously, that condition has the meaning that a massless spin (helicity)-1 particle is part of the spectrum of the theory. The level-1 analysis corroborates that interpretation.

Next, we analyse the $M = 2$ (second level) excited states. The integer 2 can be partitioned as (2,0) and (1,1). This leads to the most general level-2 excited states of the form

$$|\psi_2\rangle = \{\alpha_{\mu\nu} a_{(1)}^\mu a_{(1)}^\nu + \beta_\mu a_{(2)}^\mu\} |0, P\rangle \quad (16.78)$$

The naive degeneracy in this case, as given by the number of parameters in $|\psi_2\rangle$ is $\frac{D(D+1)}{2} + D = \frac{D(D+3)}{2}$. The mass shell condition now yields $P^2 = 1$ which is a *massive* excitation. This permits to analyse in the frame $P_\mu = (M, \mathbf{0})$, the rest frame. Now both $L_1|\psi_2\rangle = 0$ and $L_2|\psi_2\rangle = 0$ have to be imposed. This results in

$$\begin{aligned} |phys_2\rangle = & \alpha_{ij} [a_{(1)}^{i\dagger} a_{(1)}^{j\dagger} - \frac{\delta_{ij}}{D-1} \sum_{k=1}^{D-1} a_{(1)}^{k\dagger} a_{(1)}^{k\dagger}] |0, P\rangle \\ & + (\sum_{i=1}^{D-1} \alpha_{ii}) [\sum_{i=1}^{D-1} a_{(1)}^{i\dagger} a_{(1)}^{i\dagger} + \frac{D-1}{5} (a_{(1)}^{0\dagger})^2 - 2a_{(2)}^{0\dagger}] |0, P\rangle \\ & + \beta_i [a_{(2)}^{i\dagger} + a_{(1)}^{0\dagger} a_{(1)}^{i\dagger}] |0, P\rangle \end{aligned} \quad (16.79)$$

Though the degeneracy has come down, the time-component excitations are still there. In the above, i, j take on values 1 to $D-1$. The first term denotes states with positive norm with degeneracy that of a *massive spin-2* particle in $D-1$ *spatial*

dimensions. The second set are all zero-norm states which can further be represented as $L_1^\dagger a_{(1)}^i |0, P\rangle$. They are, therefore, spurious states and decouple. But the time-component excitations are still not eliminated and are present in the last term.

The last term represents a massive *spin-0* state. It should be recalled that such a state is expected from the first leading daughter trajectory, provided it has a positive norm. Elementary oscillator algebra gives its norm to be $2(D-1) + \frac{(D-1)^2}{25}(2-4) = 2\frac{(D-1)(26-D)}{25}$. For $D < 26$ it is a positive norm state and physical. For $D > 26$ it is a negative norm state which does not decouple and the theory is inconsistent. At exactly $D = 26$ it becomes a zero-norm state and only for this value of D it can also be written as $(2L_2^\dagger + 3L_1^{\dagger 2})|0, P\rangle$ and is therefore spurious and decouples.

16.5 The Shapiro-Virasoro Model

So far our entire focus has been on what one may call *Veneziano* models whose starting point was the Veneziano formula for the four-point function. Though Veneziano's original work [23] involved both vector and scalar particles ($\pi\pi \rightarrow \pi\omega$), subsequently we considered scattering of neutral spin-less particles. In this section, we shall briefly discuss alternative models developed by Shapiro and Virasoro as they are in a sense fundamentally different from the Veneziano models.

The starting point for this discussion is the Veneziano amplitude of Eq. (15.14) for the scattering of identical neutral spin-less particles:

$$A_4(s, t, u) = \bar{\beta}\{B(-\alpha(s), -\alpha(t)) + B(-\alpha(t), -\alpha(u)) + B(-\alpha(u), -\alpha(s))\} \quad (16.80)$$

where $\bar{\beta}$ is some (coupling)constant, and $B(x, y) = B(y, x)$ is the Euler Beta-function:

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \quad (16.81)$$

with $\Gamma(x)$ the Euler Gamma function. Two very important properties here are (i) precise duality displayed by each of the terms in the sense of Eq. (15.21), and, (ii) absence of all odd-daughters in each of the three terms owing to a single condition $\Sigma = \alpha(s) + \alpha(t) + \alpha(u) = -1$, which was also equal to the condition $\alpha(0) = 1$. Of course, the assumption of a linearly rising Regge trajectory is essential.

On using the properties of the Euler Gamma functions, it is easy to see that the three terms add to:

$$A_4 = \bar{\beta} \frac{\Gamma(-\frac{\alpha(s)}{2})\Gamma(-\frac{\alpha(t)}{2})\Gamma(-\frac{\alpha(u)}{2})}{\Gamma(1 + \frac{\alpha(s)}{2})\Gamma(1 + \frac{\alpha(t)}{2})\Gamma(1 + \frac{\alpha(u)}{2})} \quad (16.82)$$

This is still the Veneziano formula, recast in a different way.

When $\Sigma = -1$, this formula can be written as the sum of three terms satisfying the said properties. But Virasoro [24] showed how the formula above can be used even when $\Sigma \neq -1$! Then, it can no longer be written as the sum of three terms

each of which satisfies duality. Technically this is called *non-planar duality* as against *planar-duality* when $\Sigma = -1$. We shall not go into the meaning of this terminology here.

The obvious next issue was generalization to N -point functions. This was successfully carried out by Shapiro in [25]. The model was subsequently named the Shapiro-Virasoro(SV) model. Shapiro found that in order to obtain the N -point generalization, some drastic changes were necessary. Firstly, he noted that Σ could not be arbitrary, but had to be exactly $\Sigma = -2$. He also noted that the ground state had to be at $p^2 = -2$ as against the tachyon of the Veneziano model at $p^2 = -1$. The Shapiro-Virasoro ground state is also *tachyonic*. Together, these lead to $\alpha(0) = 2$, a condition that will be seen to have very deep repercussions.

Shapiro found that the Koba-Nielsen variables had to be doubled too, from z to $(z, \bar{z})$. We reproduce here Shapiro's expression for the N -point function of the SV-model in terms of the doubled Koba-Nielsen variables (see Eq. (49) of [7]):

$$B_N^{SV} = \int \frac{\prod_{i=1}^N d^2 z_i}{dV_{abc}} \prod_{i < j} |z_i - z_j|^{\alpha' p_i \cdot p_j} \quad (16.83)$$

with

$$dV_{abc} = \frac{d^2 z_a d^2 z_b d^2 z_c}{|z_a - z_b|^2 |z_b - z_c|^2 |z_c - z_a|^2} \quad (16.84)$$

the projective invariance is now the larger invariance under *complex projective transformations* under which $(z, \bar{z})$ no longer transform the same way.

The theme of doublings continues when an oscillator representation of the N -point functions of SV-model is sought(almost). The factorization procedure can also be generalized for the SV-model on introduction of *two* sets of commuting oscillators [18, 19]: in the metric conventions of [7] (see his Eq. (133)). However, the zero-mode oscillators are not doubled

$$[\hat{Q}^\mu, \hat{P}^\nu] = i\eta^{\mu\nu} \quad (16.85)$$

The Virasoro operators generalize to

$$\begin{aligned} Q(z, \bar{z}) = & \hat{Q} - 2\alpha' \ln z\bar{z} + i\sqrt{\frac{\alpha'}{2}} \sum_{n=1}^{\infty} \frac{a_{(n)} z^{-n} - a_{(n)}^\dagger z^n}{\sqrt{n}} \\ & + i\sqrt{\frac{\alpha'}{2}} \sum_{n=1}^{\infty} \frac{\bar{a}_{(n)} \bar{z}^{-n} - \bar{a}_{(n)}^\dagger \bar{z}^n}{\sqrt{n}} \end{aligned} \quad (16.86)$$

The (tachyon)vertex operator generalizes to

$$V(z, \bar{z}; p) = V(z; \frac{p}{2}) V(\bar{z}; \frac{p}{2}) \quad (16.87)$$

The fact that even in the SV-model there is still only one set of zero-mode oscillators necessitates the equal sharing of the momenta between the vertex operators. The operators (L_0, L_n, R) also get doubled to $(L_0, \bar{L}_0, L_m, \bar{L}_m, R, \bar{R})$ with the additional oscillators also obeying a centrally extended Virasoro algebra with the *same* central extension. The physical state conditions generalize to

$$L_m|phys\rangle = \bar{L}_m|phys\rangle = 0 \quad (L_0 - 1)|phys\rangle = (\bar{L}_0 - 1)|phys\rangle = 0 \quad (16.88)$$

Levels are now labelled by two numbers $M, \bar{M}$ which are eigenvalues of $R, \bar{R}$. However, consistency requires $M = \bar{M}$. The on-shell condition for the level $m = \bar{M}$ becomes $2 + \frac{\alpha'}{2}s = M + \bar{M} = 2M$. Thus, the level $M = \bar{M} = 1$ is a massless excitation, but with spin(helicity) equal to 2! So here too $\alpha(0) = 2$ has the same physical interpretation as $\alpha(0) = 1$ before. Such an excitation is clearly not a part of the hadronic world and in fact the Froissart bound is violated by it. But when extended to String Theories as models for dual models, this becomes a central feature as a way of integrating gravitation with other known forces!

Later it turns out that the String model appropriate to the Veneziano model is that of an *Open String*, while that underlying the Shapiro-Virasoro model is that of a *Closed String*. Furthermore, it turns out that the two cannot be separated. Unitarizing the first is impossible without including the second!

References

1. S. Fubini, G. Veneziano, Level structure of dual resonance models. *Nuovo Cimento* **64A**, 811 (1969)
2. S. Fubini, D. Gordon, G. Veneziano, *Phys. Lett. B* **29**, 679 (1969)
3. L. Susskind, *Phys. Rev. Lett.* **23**, 545 (1969)
4. L. Susskind, *Phys. Rev. D* **1**, 1182 (1970)
5. E. Galli, L. Susskind, *Phys. Rev. D* **1**, 1189 (1970)
6. Y. Nambu, Quark model and factorisation, in *International Conference On Symmetries and Quark Model*, ed. by R. Chand (Gordon and Breach, 1970)
7. P. Di Vecchia, The birth of string theory, in *String Theory and Fundamental Interactions*, Lecture Notes on Physics, vol. 737 (2008), pp. 59–118. [arXiv:1704.0101](https://arxiv.org/abs/1704.0101)
8. P.H. Frampton, *Dual Resonance Models* (Benjamin, W.A, 1974)
9. B. de Wit, J. Smith, *Field Theory in Particle Physics Vol I* (North Holland Personal Library, 1986)
10. A. Peres, *Quantum Theory: Concepts and Methods* (Kluwer Academic Publishers, 1993)
11. D.B. Fairlie, The birth of string theory, in *String Theory and Fundamental Interactions*, Lecture Notes on Physics, vol. 737
12. S. Fubini, G. Veneziano, *Nuovo Cimento* **67**, 29 (1970)
13. G.H. Hardy, S. Ramanujan, *Proc. Lond. Math. Soc.* **17**, 75 (1917)
14. R. Hagedorn, *Nuovo Cimento (Supp)* **3**, 147 (1965)
15. S. Fubini, G. Veneziano, *Ann. Phys.* **63**, 12 (1971)
16. L. Clavelli, J.H. Weis, (1970) (unpublished)
17. M.A. Virasoro, *Phys. Rev. D* **1**, 2933 (1970)
18. E. Del Giudice, P. Di Vecchia, *Nuovo Cimento* **A5**, 90 (1971)
19. M. Yoshimura, *Phys. Lett. B* **34**, 79 (1971)
20. E. Del Giudice, P. Di Vecchia, S. Fubini, *Ann. Phys.* **70**, 378 (1972)

- 21. R.C. Brower, Phys. Rev. D **6**, 1655 (1972)
- 22. P. Goddard, C.B. Thorn, Phys. Lett. B **40**, 235 (1972)
- 23. G. Veneziano, Nuovo Cimento **57A**, 190 (1968)
- 24. M.A. Virasoro, Phys. Rev. **177**, 2309 (1969)
- 25. J.A. Shapiro, Phys. Lett. B **33**, 361 (1970)



17.1 Introduction

This chapter describes the genesis of string theory. The dual resonance models and the S-matrix programme that we have discussed so extensively played dedicated midwives to this extraordinary birth! Poincare was among the earliest to think of theories of strings which he called *Threads* [1]!

The development of string theory that we will be concerned within this chapter is due to the more or less simultaneous breakthroughs by Nambu [2,3], Susskind [4–6], and, Holger Bech Nielsen [7]. The motivations of Nambu as well as that of Susskind were somewhat closer, both being driven by the oscillator structures underlying dual resonance models. While Nambu highlighted the oscillator structure to display *factorizability* of the dual amplitudes, Susskind was seeking to obtain the dual resonance models themselves starting from oscillators. Nielsen's approach was completely different, as he was trying to understand Feynman diagrams of very high order. In the process he came to recognize the importance and dominance of a certain class of diagrams which he called *fishnet* diagrams, which were more precisely named *planar diagrams* later on. Remarkably, Nielsen saw a natural interpretation of the planar diagrams in terms of a world-sheet of a string swept out during its time evolution. We refer the reader to Nielsen's historical account of the genesis of String Theory [8].

In Sect.4 of Susskind's paper he explicitly introduces a *Rubber String Model* as a *continuum* description of his dual model oscillators. He goes on to develop what came to be known later as the World-sheet formalism for these strings. In Nielsen's approach, the world-sheet was realized by the planar diagrams for very large order Feynman diagrams for scattering of quarks. In that picture, Nielsen envisaged momentum flows in and out of the world-sheet as the exchange of momenta with the quarks. In an important step, Nielsen had also proposed an action for the strings.

In Sects. 10 and 11 of his paper, Susskind provides a detailed map between his and Nielsen's ideas and establishes their equivalence.

We shall mostly follow Nambu's line of thinking here as it provides a clear and straightforward exposition of the ideas, as augmented by the ideas of Goto [9]. The action for the string theory that played a central role in subsequent developments came to be known as the *Nambu-Goto Action*. It will play an equally central role in our descriptions of *Effective String Theories*.

17.2 Emergent Strings and Their Actions

17.2.1 Nielsen-Susskind Action

The infinitely many oscillators underlying the dual resonance models naturally led to considerations of field theories or theories of *extended objects* as possibly deeper descriptions of the dual models themselves. This is much in the same spirit as Sect. 6.2 of our Chap. 6, where we constructed a field theory from the creation and annihilation operators of particle states.

There is, however, a very important difference of detail, which we wish to elaborate on. In the free-particle case, the oscillators were labelled by the continuous variables $\mathbf{k}$, the momenta of the particles. That constituted the *mode expansion* in that case. But the oscillators in the dual models, for both the Veneziano-like models as well as the Shapiro-Virasoro type models, are labelled by an *integer*, although of infinite range. That raises two possibilities for the types of field theories to be considered: (a) a one spatial-dimensional field theory with boundaries and appropriate boundary conditions, and, (b) where the spatial topology is that of a circle.

Even though at this point there is nothing that specifically suggests a consideration of strings, we shall shortly see that case (a) will lead to a theory of open-strings, while case (b) to that of closed strings. Eventually one will also find a deep connection between the two. The reader should recall our elaborate treatment of the dual resonance models where too these physically distinct possibilities were foreseen, i.e. open-strings description for the Veneziano type models and closed-string descriptions for the Shapiro-Virasoro models. It is also worth emphasizing that the Koba-Nielsen variables, living on a circle, fit in with discretely labelled (by integer) modes rather naturally. The reader should also refer to the discussions in [10, 11].

The picture of an elastic string as what underlies the dual resonance amplitudes was pretty much explicit in both Nambu's as well as Nielsen's approaches. Picturing the elastic string to be the limit of a linear chain of a very large number N of mass-points with each mass-point interacting with its neighbours through a harmonic potential, Nambu first proposed the action

$$S = \int \int d\xi d\tau \left(\frac{\partial X^\mu}{\partial \tau} \frac{\partial X_\mu}{\partial \tau} - \frac{\partial X^\mu}{\partial \xi} \frac{\partial X_\mu}{\partial \xi} \right) \quad (17.1)$$

to describe a *Relativistic String*. As can be seen from the above mentioned analyses in Susskind's paper, this is the same action that had been considered by both Nielsen

and Susskind. We shall henceforth call this the *Nielsen-Susskind String Action* to distinguish it from the *Nambu-Goto String action* to be discussed shortly.

There are many features of this action that need an explanation. ξ, τ are the so-called *World-Sheet* coordinates. To understand these, recall that each mass-point traces out a *World-line* which is nothing but its trajectory in space-time. Such a trajectory can certainly be described by specifying the spatial coordinates as a function of time $X^i(t)$. To make a relativistic description one introduces $X^\mu = X^i, t$. Alternatively, one can introduce the World-line description, $X^\mu(\tau)$ where the space-time coordinates of a point on the trajectory are given in terms of the *proper-time* of the particle. Now each of the N mass points has its own world-line and together they span the world-sheet. The world-sheet coordinate τ need not necessarily have any interpretation as a proper-time, but the sheet must be a Minkowskian two-dimensional surface. This explains the negative sign in the above equation.

To further gain some understanding of this action proposed by Nambu, Nielsen and Susskind, let us quickly revisit a classical string. We will treat this also as the limiting description of a very large number N of non-relativistic particles, interacting with only the nearest neighbour through a harmonic oscillator potential (see my lecture notes [12]). We will take space to be one-dimensional too to simplify the presentation, although it is not essential. The reader is urged to generalize the treatment given. The Lagrangean for the system, with the further simplifying assumption of equal mass for each particle is

$$L = \frac{1}{2} m \sum_i \dot{X}_i^2 - \frac{k}{2} \sum_i (X_{i+1} - X_i)^2 \quad (17.2)$$

Here $X_i(t)$ is the position of the mass-point labelled i . The system is *translationally invariant*, i.e. the Lagrangean is invariant under $X_i \rightarrow X_i + b$ with b being any constant. Additionally, let us assume the end-point masses are held fixed, i.e. $X_1(t) = b_1$; $X_N(t) = b_N$. The equilibrium positions $X_i^{(0)}$ characterized by $\dot{X}_i^{(0)} = \text{const}$ are given by $(X_{i+1}^{(0)} - X_i^{(0)}) - (X_i^{(0)} - X_{i-1}^{(0)}) = 0$ with the explicit solutions being given by $X_i^{(0)} = X_1^{(0)} + i \cdot a$. Using the translational invariance of the system, one can fix $X_1(t) = 0$. Then $X_N(t) = (N-1)a$. The separation of the point masses at equilibrium is uniform and equal to a .

Now we pass to the limit of the number of mass-points N tending to infinity but keeping the translationally invariant “length” of the string $l = (N-1)a$ fixed. This would require that the “spacing” a should tend to zero. Then $\sum_i \rightarrow \int \frac{dx}{a}$, and the Lagrangean tends to

$$L \rightarrow \frac{m}{2} \int \frac{dx}{a} (\dot{X}(x, t)^2) - \frac{k}{2} \int \frac{dx}{a} \left(\frac{X(x+a, t) - X(x, t)}{a} \right)^2 a^2 \quad (17.3)$$

Therefore, for the limit to make sense, one must also demand that $m \rightarrow 0$ in such a way that $\frac{m}{a} \rightarrow \mu$ and also $ka \rightarrow Y$. The Lagrangean then becomes

$$L = \int dx \left\{ \frac{1}{2} \mu \dot{X}^2 - \frac{1}{2} Y X'^2 \right\} \quad (17.4)$$

where $\dot{X}$, X' stand for the temporal and spatial derivatives, respectively. In terms of an action, this is

$$S = \int \int dx dt \left\{ \frac{\mu}{2} \dot{X}^2 - \frac{Y}{2} X'^2 \right\} \quad (17.5)$$

μ has the significance of the mass per unit length while Y has the significance of the Young's modulus. If the elasticity of the medium is *isotropic*, this action can be generalized to more than one spatial dimensions as:

$$S = Y \int \int dx dt \left\{ \frac{\mu}{2Y} \dot{X}_i^2 - \frac{1}{2} X_i'^2 \right\} \quad (17.6)$$

The index now goes over the $(D - 1)$ spatial directions only.

Despite looking very similar to Nambu's action of Eq. (17.1), there are profound differences. Of course, it is interesting that the action for this non-relativistic string theory almost looks relativistic. It too has the look of a world-sheet action with (x, t) acting like world-sheet coordinates, even with a correct Minkowski signature. The harmonic oscillator interaction between the point masses, which can be taken as defining the elastic string, was central to that. Let us list three crucial differences: (i) the world-sheet coordinates for the non-relativistic string are seen to be fixed to be (x, t) , which are part of the coordinates of the D-dimensional space-time itself. This is in contrast to the action of Eq. (17.1) where these are (ξ, τ) with no obvious connection to the D-dimensional coordinates, (ii) while the points of the world-sheet in Nambu's case are labelled by the fields $X^\mu(\sigma, \tau)$, the non-relativistic strings of Eq. (17.6) are labelled only by the spatial $X^i(x, t)$. The difference is of course due to relativistic invariance, much as in the case of the relativistic particle. But there X^0 could be identified with time, but now there is a field $X^0(\sigma, \tau)$ for each value of σ , and it is not as yet obvious that it can everywhere be identified with time, (iii) in place of the universal velocity of light (chosen to be unity) of the relativistic action, the non-relativistic action explicitly involves the non-universal velocity $\sqrt{\frac{Y}{\mu}}$ which is the velocity of propagation of the elastic string waves. It should be mentioned at this point that there should be a dimensional factor in front of the action of Nambu and finally, (iv) while the non-relativistic string admits both longitudinal and transverse excitations, the Nielsen-Susskind action admits only transverse excitations.¹

It will soon be seen that all these issues are deeply interconnected. Underlying these interconnections are the invariances of the action of Eq. (17.1); Nambu mentions only the invariances under $\tau \rightarrow \tau + c$, $X^\mu \rightarrow X^\mu + a^\mu$, but it turns out that the invariances of this action are *infinite-dimensional*.

¹ The higher dimensional generalization of the non-relativistic string is somewhat unrealistic as it predicts equal velocities for transverse and longitudinal waves. A more realistic treatment of elastic media, as for example the one in Landau and Lifshitz book [13] (see their Sect. 22), always yields

$c_l > \sqrt{\frac{4}{3}} c_t$.

17.2.2 Nambu-Goto Action

Nambu [3], and independently Goto [9] proposed the action

$$S_{NG} = -\frac{1}{2\pi\alpha'} \int \int d^2\xi \sqrt{-\det g} \quad g_{\alpha\beta} \equiv \frac{\partial X^\mu}{\partial \xi^\beta} \frac{\partial X_\mu}{\partial \xi^\alpha} \quad (17.7)$$

Following standard practice, we have set $\hbar = c = 1$. The world-sheet coordinates are now ξ^1, ξ^2 . This is the *area* of the world-sheet and is endowed with an obviously *geometrical* meaning. This geometrical aspect of this action was the main motivation for Nambu to introduce this (though he prefaces his introducing this with the words *For curiosity, then, let us try to construct a geometric action integral...*).

As a geometric entity, it should remain unchanged under general reparametrizations of ξ^α . To that extent the situation is reminiscent of Einstein-Hilbert general relativity theory. But there is a profound difference that is often unstated in literature; while in general relativity the changes to the action from general coordinate transformations are exactly compensated by the changes due to the transformation of the intrinsic metric of the manifold, in the case of the Nambu-Goto action, general coordinate transformations of the world-sheet coordinates leaves the action invariant even though there is no *intrinsic world-sheet metric*! In that sense, the general coordinate invariances of General Relativity Theory are devoid of physical significance, as Kretschmann [14] emphasized soon after GR was formulated. We will return to this important aspect several times in the book.

Nambu says that what we called Nielsen-Susskind action is not a purely geometrical quantity in contrast to the Nambu-Goto action. Di Vecchia in [11] also criticizes the Nielsen-Susskind action (see the remarks immediately following his Eq. (260) of his Sect. 9). In particular he claims that there was no reason to expect the Virasoro gauge conditions, and also why they should vanish classically. We wish to point out that the two actions are intimately related, somewhat the way mothers and daughters are! Though it is true that the Nambu-Goto action is invariant under arbitrary world-sheet coordinate transformations while the Nielsen-Susskind action is only invariant under conformal transformations, the latter is a *gauge-fixed* form of the former after a so-called conformal gauge is chosen, and the residual coordinate invariances of this gauge-fixed action are precisely the conformal transformations. This is discussed at length in our papers [15, 16]. Therefore, the Nielsen-Susskind action must always be used in conjunction with the conformal gauge conditions. The latter are precisely the classically vanishing conditions alluded to by Di Vecchia. Later on, we shall give an alternate approach to such actions based on the Effective String approach of Polchinski and Strominger [17].

17.3 Classical Analysis of the Nambu-Goto Action

Our ultimate goal is to understand the quantum behaviour of strings as described by the Nambu-Goto action, and to see to what extent they can explain the dual resonance models. The major development was the work of Goddard, Goldstone, Rebbi and Thorne (GGRT) [18] who laid the groundwork for the quantization of relativistic string theories. We wish to first bring out the classical aspects in preparation for the eventual quantization. For both the classical and quantum aspects, we wish to draw heavily from the analyses of [3, 9, 11, 18]. The four sources emphasize different nuances and we have tried to blend all four of them!

A particularly popular method of quantization is to first study the so-called normal-modes classically, and then promote the coefficients of the mode-expansion to quantum-mechanical operators. This is particularly so when the system in question is *free*, in a sense to be made more precise as we go along. The textbook cases of the quantization of the free scalar, Dirac, and Maxwell's (electromagnetic field) follow this route. The reader is referred to the excellent books on the subject [19–21].

Trouble arises with this approach when the equations of motion do not admit *unique* solutions. How can that be, one may wonder. In fact, this difficulty manifests both in the quantization of the Nambu-Goto string theory, as well as the much more familiar (and easier looking) Quantum Electrodynamics. We proceed to explain the crux of the difficulty in the case of QED. The equations of motion of *classical* electrodynamics, when formulated in terms of the *Vector Potentials* is

$$\partial_\mu F^{\mu\nu} = \partial^\nu \partial \cdot A - \square A^\nu = 0 \quad (17.8)$$

This equation cannot have unique solutions because if A^ν is a solution, so is its *gauge transform* $A^\nu + \partial^\nu \Lambda(x)$, for arbitrary $\Lambda(x)$. This arises due to the invariance of the electromagnetic action (and Lagrangean density) under gauge transformations. A way to get around this difficulty is to *fix* the gauge arbitrariness by requiring, for example, $\partial \cdot A = 0$. Then the gauge-fixed equation of motion becomes

$$\square A^\nu = 0 \quad (17.9)$$

for which solutions can indeed be found. They are the plane-wave solutions $\epsilon^\nu e^{-ik \cdot x}$. The general solution can be expanded in terms of the modes (see, for example, Sect. 3.2 of [22]); the reader should however note differences in our approach.

$$A^\mu(x) = \int \frac{d^3k}{2k_0(2\pi)^3} (a^\mu(k)e^{-ik \cdot x} + a^{\mu\dagger}(k)e^{ik \cdot x}) \quad (17.10)$$

Usually, the mode expansion for the photon field explicitly displays the polarization vectors, which are two for physical states, and four as a starting point in a covariant treatment. While the correct number of polarizations for physical states of the electromagnetic field, which amounts to the important step of determining the physical states of a theory, is going to play a central role in the ensuing discussion of string

quantization, we shall not dwell further on it in this electromagnetic context, as we wish to highlight other aspects. To those who wish to connect with the polarization vectors, it should be noted that we have clubbed $\sum_{\lambda} a^{(\lambda)}(k) \epsilon_{(\lambda)}^{\mu}(k)$ of [22] into the single factor $a^{\mu}(k)$.

On passage to quantum theory, these mode expansion coefficients become operators, the usual creation and annihilation operators. In the conventions of Itzykson and Zuber [22], the mode expansion coefficients will obey the quantum commutation relations

$$[a^{\mu}(k), a^{\nu\dagger}(k')] = -g^{\mu\nu} 2k_0 (2\pi)^3 \delta^{(3)}(\mathbf{k} - \mathbf{k}') \quad (17.11)$$

In contrast, the canonical commutation relations of [22] read:

$$[a^{(\lambda)}(k), a^{(\lambda')\dagger}(k')] = -g^{\lambda\lambda'} 2k_0 (2\pi)^3 \delta^{(3)}(\mathbf{k} - \mathbf{k}') \quad (17.12)$$

Consequently, in the quantum theory, $\partial \cdot A$ itself becomes an operator:

$$\begin{aligned} \partial \cdot A &= -i \int \frac{d^3 k}{2k_0 (2\pi)^3} \{k \cdot a(k) e^{-ik \cdot x} - k \cdot a^{\dagger}(k) e^{ik \cdot x}\} \\ &= \partial \cdot A^{(+)} + \partial \cdot A^{(-)} \end{aligned} \quad (17.13)$$

where the superscripts $\pm$ refer to the positive and negative frequency parts of $\partial \cdot A$ respectively. More precisely, the $+$ part is the one involving $a(k)$, and the negative part the one involving $a^{\dagger}(k)$. It is easy to check that $[\partial \cdot A^{(+)}(x), \partial \cdot A^{(-)}(x')] = C(x - x') \neq 0$. Here $C(x - x')$ is a so-called c-number and not an operator. Consequently, $[\partial \cdot A(x), \partial \cdot A(x')]$ is also a non-vanishing c-number.

Therefore, in the quantum theory it is impossible to impose this as a constraint on physical states, i.e. $\partial \cdot A(x)|phys\rangle = 0$ everywhere. The best that can be done is to impose it *weakly* on the physical states, i.e. $\langle \psi_1 | \partial \cdot A(x) | \psi_2 \rangle = 0$. This can equally well be implemented by requiring only the positive-frequency parts of this constraint to vanish on physical states, i.e. $(\partial \cdot A)^{(+)}|phys\rangle = 0$.

Another subtle point is that even after fixing the gauge (for the moment, let us go back to the classical description) by $\partial \cdot A = 0$, the gauge is not fully fixed as further gauge transformations of the type $\partial^{\mu} \chi$ with $\square \chi = 0$ are still possible. In electrodynamics this additional freedom called *Residual Gauge Transformations* is harmless, but in the Nambu-Goto string theory, as also for example, in General Relativity theory, this becomes very important.

The invariances that lead to all these technicalities in the case of Nambu-Goto theory is it's invariance under arbitrary transformations among the world-sheet coordinates which induce the transformations $\delta X^{\mu} = -\xi \cdot \partial X^{\mu}$ among X^{μ} ; here ξ^{α} are the parameters of infinitesimal world-sheet coordinates. Often, this freedom is fixed through the coordinate conditions $\dot{X} \cdot X' = 0$; $\dot{X}^2 + X'^2 = 0$, where dot refers to derivatives wrt τ and X' to derivatives wrt σ . These conditions are fixed because of their inherent geometric meanings as fixing a locally orthonormal system of coordinates on the world-sheet. They are analogous to the $\partial \cdot A = 0$ gauge-fixing conditions of Electrodynamics that we just discussed. In both cases, the gauge(coordinate)-fixing conditions themselves are “covariant”.

Writing the action in the familiar form of $S_{NG} \equiv \int d\tau \int d\sigma \mathcal{L}$ (this is just a change in notation as far as the world-sheet coordinates are concerned, they have been renamed (σ, τ) , nothing more), the principle of least-action, with the end-points fixed, $\delta X^\mu(\sigma, \tau_i) = \delta X^\mu(\sigma, \tau_f) = 0$, would give the classical equations of motion, the Euler-Lagrange equations, as:

$$\frac{\partial}{\partial \tau} \frac{\partial \mathcal{L}}{\partial \dot{X}^\mu} + \frac{\partial}{\partial \sigma} \frac{\partial \mathcal{L}}{\partial X'^\mu} = 0 \quad (17.14)$$

The variational principle includes the boundary contributions too, if any. So, at this stage one has to specify if we are dealing with an open-string or a closed string. The latter is defined by $X^\mu(\sigma = 0, \tau) = X^\mu(\sigma = \pi, \tau)$. There will be no boundary conditions coming from the action principle. The open string case, on the other hand, requires the boundary conditions

$$\{\delta X^\mu \frac{\partial \mathcal{L}}{\partial X'^\mu}\}_{\sigma=0}^{\sigma=\pi} = 0 \quad (17.15)$$

The end-points of σ -space have been taken to be $(0, \pi)$ without loss of generality. At this point, in principle, four possibilities arise: (i) $\delta X^\mu(0, \tau) = \delta X^\mu(\pi, \tau) = 0$, (ii) $\frac{\partial \mathcal{L}}{\partial X'^\mu}(0, \tau) = \frac{\partial \mathcal{L}}{\partial X'^\mu}(\pi, \tau) = 0$, (iii) $\delta X^\mu(0, \tau) = \frac{\partial \mathcal{L}}{\partial X'^\mu}(\pi, \tau) = 0$ and finally, (iv) $\frac{\partial \mathcal{L}}{\partial X'^\mu}(\pi, \tau) = 0$. The boundary conditions of the type $\delta X^\mu = 0$ are called the *Dirichlet* boundary conditions, while the other type, in the present context would be *Neumann* boundary conditions. Then, (i) are pure Dirichlet, (ii) pure Neumann, and the other two mixed Dirichlet-Neumann type. Goddard et al only considered the pure Neumann-type ii. In fact Dirichlet conditions were not invoked in string theory for a long time, till Polchinski used them to dramatic effect in his discovery of D-branes [23].

Before carrying out any explicit calculations, it is worthwhile to bring out additional invariances of the Nambu-Goto action. These have to do with the fact that in the Minkowski space labelled by X^μ (called Target Space by String theorists), one has the Poincare group of invariances consisting of translations $\delta X^\mu = a^\mu$ (constant vector), and, Lorentz transformations $\delta X^\mu = \Lambda^\mu_\nu X^\nu$. These lead to associated conservation laws. The translational invariance leads to the conservation of momentum:

$$\frac{\partial}{\partial \tau} \mathcal{P}_\tau^\mu + \frac{\partial}{\partial \sigma} \mathcal{P}_\sigma^\mu = 0 \quad \mathcal{P}_\tau^\mu = \frac{\partial \mathcal{L}}{\partial \dot{X}_\mu} \quad \mathcal{P}_\sigma^\mu = \frac{\partial \mathcal{L}}{\partial X'^\mu} \quad (17.16)$$

The *energy-momentum-currents* $\mathcal{P}^\mu$ are defined by

$$\mathcal{P}_\tau^\mu \equiv \frac{\partial \mathcal{L}}{\partial \dot{X}_\mu} \quad \mathcal{P}_\sigma^\mu \equiv \frac{\partial \mathcal{L}}{\partial X'^\mu} \quad (17.17)$$

It should be noted that Eq. (17.16) is also the Euler-Lagrange equations of motion Eq. (17.14). $(\mathcal{P}_\tau^\mu, \mathcal{P}_\sigma^\mu)$ can be thought as the components of *Energy-Momentum Current*. By standard manipulations it can be seen that

$$P^\mu = \int_0^\pi d\sigma \mathcal{P}_\tau^\mu \quad (17.18)$$

is the conserved momentum. Likewise, invariance under the Lorentz transformations leads to the conserved generators $M^{\mu\nu}$. We now display the explicit expressions for the $\mathcal{P}^\mu$ with $T_0 \equiv \frac{1}{2\pi\alpha'}$:

$$\mathcal{P}_\tau^\mu(\sigma, \tau) = -T_0 \frac{(X' \cdot \dot{X})X'^\mu - X'^2 \dot{X}^\mu}{\sqrt{(X' \cdot \dot{X})^2 - X'^2 \dot{X}^2}} \quad \mathcal{P}_\sigma^\mu = -T_0 \frac{(X' \cdot \dot{X})\dot{X}^\mu - X'^2 X'^\mu}{\sqrt{(X' \cdot \dot{X})^2 - X'^2 \dot{X}^2}} \quad (17.19)$$

Goddard et al bring out some very simple physical interpretations underlying the above equations, without even solving them. To do so, they consider a Lorentz frame in which $\tau = t$. Then the evolution of the world-sheet is described by $\mathbf{x}(\sigma, t)$. Then they introduce the vector $d\mathbf{s} = \frac{\partial \mathbf{x}}{\partial \sigma}$ along the string. The *transverse velocity* $\mathbf{v}_\perp$ is then given by

$$\mathbf{v}_\perp = \frac{\partial \mathbf{x}}{\partial t} - \frac{\partial \mathbf{x}}{\partial s} \left(\frac{\partial \mathbf{x}}{\partial t} \cdot \frac{\partial \mathbf{x}}{\partial s} \right) \quad (17.20)$$

It is straightforward to show $d\mathbf{s} \cdot \mathbf{v}_\perp = 0$ (Hint: $\frac{\partial \mathbf{x}}{\partial s}$ is a unit-vector.) After some tedious, but not difficult algebra, the Nambu-Goto action can be recast as

$$S_{NG} = -T_0 \int_{t_i}^{t_f} dt \int_0^\pi d\sigma \frac{ds}{d\sigma} \sqrt{(1 - \mathbf{v}_\perp^2)} \quad (17.21)$$

whose interpretation as the sum of relativistic point-particle actions is immediate.

At this point GGRT choose σ so as to satisfy $\frac{\partial \mathbf{x}}{\partial t} \cdot \frac{\partial \mathbf{x}}{\partial \sigma} = 0$ (the choice of $\tau = t$ still continues). Geometrically this choice ensures that the path of a point with fixed σ is perpendicular to the string. Together with the choice $\tau = t$ this is equivalent to $\dot{X} \cdot X' = 0$. The first simplification is that the velocity $\frac{\partial \mathbf{x}}{\partial t}$ is now entirely transverse, i.e. it equals $\mathbf{v}_\perp$. The $\mathcal{P}^\mu$ also simplifies to:

$$\mathcal{P}_\tau^\mu = T_0 \sqrt{\frac{-X'^2}{\dot{X}^2}} \dot{X}^\mu \quad \mathcal{P}_\sigma^\mu = T_0 \sqrt{\frac{-\dot{X}^2}{X'^2}} X'^\mu \quad (17.22)$$

The boundary conditions, requiring the vanishing of $\mathcal{P}_\sigma^\mu$ at $(0, \pi)$ immediately imply that the ends of the string move with $v_\perp = c$! In between, on using $X'^2 = (\frac{ds}{d\sigma})^2$, it is easy to show that $\sqrt{-X'^2 \dot{X}^2} = \frac{ds}{d\sigma} \sqrt{1 - v_\perp^2}$. Putting all these together, the Euler-Lagrange equations of motion take the form

$$\frac{\partial}{\partial t} \left\{ \frac{\mathbf{v}_\perp}{\sqrt{1 - v_\perp^2}} \frac{ds}{d\sigma} \right\} = \frac{\partial}{\partial \sigma} \left\{ \sqrt{1 - v_\perp^2} \frac{\partial \mathbf{x}}{\partial s} \right\} \quad (17.23)$$

This shows that only *transverse* degrees of freedom of the string are physical. This is one of the central results of the theory. Though this has been proved with particular choice of coordinates, it will be shown subsequently that it is independent of the choice of coordinates. So far, the other coordinate condition, i.e. $X'^2 + \dot{X}^2 = 0$ has not been invoked.

GGRT, in order to bring out the nuances of both the coordinate conditions, start afresh, giving up their earlier choice of $\tau = t$. This choice can be given a pseudo-covariant guise by rewriting it as $n \cdot X = \tau$ with $n = 1, 0, 0, 0$ (in $D=4$, with obvious generalizations to higher D) and $n^2 = -1$. Now they seek a somewhat more general choice by setting

$$n \cdot X = 2(n \cdot P)\tau \quad (17.24)$$

The only requirement on n^μ they make is $n^2 \leq 0$. Towards the end, they choose $n^2 = 0$ because of the many simplifications it brings. The factor $(n \cdot P)$ with P^μ now the conserved string momentum is also for simplifying matters but is otherwise not a fundamental requirement. Likewise, for σ , they propose

$$(n \cdot P)\sigma = \pi \int_0^\sigma d\sigma' n \cdot \mathcal{P}_\tau(\sigma', \tau) \quad (17.25)$$

for fixed τ . This choice is not particularly transparent, but it does not matter as long as the σ -coordinate is fixed without ambiguity. That is in fact a general feature of gauge and coordinate fixing. Differentiating this wrt σ leads to $n \cdot \mathcal{P}_\tau = \frac{n \cdot P}{\pi}$. On combining this with the Eq. (17.16), it follows that

$$\frac{\partial n \cdot \mathcal{P}_\sigma}{\partial \sigma} = 0 \quad (17.26)$$

But the boundary conditions require $n \cdot \mathcal{P}_\sigma = 0$ at both $\sigma = 0, \pi$. Consequently, $n \cdot \mathcal{P}_\sigma = 0$ everywhere. From Eq. (17.19)

$$n \cdot \mathcal{P}_\sigma = -T_0 \frac{(n \cdot \dot{X})(\dot{X} \cdot X') - (n \cdot X')\dot{X}^2}{\sqrt{(\dot{X} \cdot X')^2 - \dot{X}^2 X'^2}} \quad (17.27)$$

But $n \cdot X' = 0$; one therefore concludes $\dot{X} \cdot X' = 0$ for this general class of coordinate conditions too, independent of further details of n^μ . On substituting this in the general expression for $\mathcal{P}_\tau^\mu$ and using $n \cdot \mathcal{P}_\tau = \frac{n \cdot P}{\pi}$ along with $n \cdot \dot{X} = 2(n \cdot P)$, it follows that $\dot{X}^2 + X'^2 = 0$. Thus the present way of fixing the coordinates not only reproduces the geometric constraints (by construction), and it also fixes the coordinates completely.

In these coordinates (with the choice $\alpha' = 1$ or equivalently $T_0 = \frac{1}{2\pi}$), the $\mathcal{P}^\mu$ take on particularly simple forms:

$$\mathcal{P}_\tau^\mu = \frac{1}{2\pi} \frac{\partial X^\mu}{\partial \tau} \quad \mathcal{P}_\sigma^\mu = -\frac{1}{2\pi} \frac{\partial X^\mu}{\partial \sigma} \quad (17.28)$$

The Euler-Lagrange eqns. also become very simple:

$$\left(\frac{\partial^2}{\partial \tau^2} - \frac{\partial^2}{\partial \sigma^2} \right) X^\mu = 0 \quad (17.29)$$

and the boundary conditions

$$\frac{\partial X^\mu}{\partial \sigma} = 0 \quad \sigma = 0, \pi \quad (17.30)$$

With the exception of the conditions $n \cdot X = 2(n \cdot P)\tau$, $(n \cdot P)\sigma = \pi \int_0^\sigma d\sigma' n \cdot P_{\tau'}$, all other equations are Lorentz-covariant and can be chosen as the basis for a *covariant quantization* of the Nambu-Goto string. We shall return to it later.

GGRT give an exceptional analysis of both a non-covariant as well as a covariant approach to quantization. As is well known, both these approaches require a *classical canonical formalism* as a stepping stone to the eventual quantization (unless one is following the Path integral approach). At the most naive level, this consists of identifying canonical coordinates and their canonical momenta, and then impose the *Canonical Poisson Brackets*, which are eventually raised to the quantum conditions as commutators. In the present context this presents several conceptual as well as technical issues. Among the conceptual ones the first is the issue of a proper interpretation of $X^0(\sigma, \tau)$; with its obvious connections to time, should this also be considered as a canonical coordinate? Not doing so would obviously impact the manifest Lorentz covariance properties. The coordinate conditions will play a fundamental role in a proper resolution of this issue.

On the technical side, the very same coordinate conditions also induce a dependence between the canonical coordinates and momenta, while the standard canonical formalism works on the premise that they are independent. To begin with, let us recast the coordinate conditions in terms of the coordinates X^μ and their canonical momenta $\mathcal{P}_\tau^\mu$ (henceforth we shall use the abbreviated notation $\mathcal{P}^\mu$ for them):

$$X' \cdot \mathcal{P} = 0 \quad \mathcal{P}^2 + \frac{1}{(2\pi)^2} X'^2 = 0 \quad (17.31)$$

The canonical formalism for systems with such constraints is well understood, and there are many paths to it. Dirac prescribed powerful techniques including those of the so-called *Dirac Brackets* [24]. The essential idea of Dirac consists of evaluating all the Poisson Brackets naively first, and then imposing the constraints. An alternate approach is to explicitly solve all the constraints first to arrive at the independent d.o.f and then perform a canonical analysis with only those. GGRT follow the second approach for the non-covariant quantization (solving the constraints explicitly is at the cost of manifest covariance). For the covariant quantization, they follow Dirac's method in spirit (but without explicitly using Dirac Brackets). GGRT show how the constraints can be explicitly solved for by recourse to the so-called *light-cone variables* for a non-covariant quantization.

17.3.1 Light-Cone Parametrization

The idea is to choose n^μ to be *light-like*, i.e. $n^2 = 0$. GGRT introduce the notation $n_\pm = \frac{n^0 \pm n^3}{\sqrt{2}}$. GGRT had $D = 4$ in mind, but we shall (so did they) treat arbitrary

D, in which case, $n_{\pm} = \frac{n^0 \pm n^{D-1}}{\sqrt{2}}$. They go further and make the explicit choice $n = \frac{1}{\sqrt{2}} (1, \mathbf{0}, -1)$. In other words, $n_- = 1, n_+ = 0, \mathbf{n} = \mathbf{0}$. Using this in the earlier equations gives $X_+ = 2P_+\tau$ and $\mathcal{P}_+ = \frac{P_+}{\pi}$. On noting that the first of these gives $X'_+ = 0$, the constraints take the form:

$$\frac{P_+}{\pi} X'_- = \mathbf{X}'_{\perp} \cdot \mathcal{P}_{\perp} \quad 2\frac{P_+}{\pi} \mathcal{P}_- = \mathcal{P}_{\perp}^2 + \frac{1}{(2\pi)^2} \mathbf{X}_{\perp}^2 \quad (17.32)$$

The meaning of these expressions is deep! The truly independent variables are the transverse d.o.f $\mathbf{X}_{\perp}$, all the rest, $X_{\pm}, \mathcal{P}_{\pm}$ are dependent variables which can be explicitly determined in terms of the independent variables through the relations above! At this stage GGRT introduce a *barycentric coordinate*:

$$q_-(\tau) = \frac{1}{\pi} \int_0^{\pi} d\sigma X_-(\sigma, \tau) \quad (17.33)$$

They then seemingly integrate the X'_- eqn above to get their Eq. (24a). We do not quite understand this eqn. whose sole purpose appears to be to get their Eq. (27d). We shall instead show how to bypass their Eq. (24a) and yet obtain their Eq. (27d) directly:

$$\dot{q}_-(\tau) = \frac{1}{\pi} \int_0^{\pi} d\sigma \dot{X}_- = 2 \int_0^{\pi} d\sigma \mathcal{P}_- \quad (17.34)$$

On using the constraint eqn. for $\mathcal{P}_-$ it is easily seen that

$$\dot{q}_-(\tau) = 2P_- = \frac{H}{P_+} \quad (17.35)$$

where use has been made of the GGRT Hamiltonian (the one that reproduces Hamilton's eqns. of motion)

$$H = 2P_+P_- = \pi \int_0^{\pi} d\sigma \left\{ \mathcal{P}_{\perp}^2 + \frac{X_{\perp}^2}{(2\pi)^2} \right\} \quad (17.36)$$

Since both H, P_+ are constants of motion, one can integrate the above to get

$$q_- = q_{0-} + \frac{H}{P_+} \tau \quad (17.37)$$

with q_{0-} also a constant of motion. It is natural then to assume

$$\{q_{0-}, P_+\} = -1 \quad (17.38)$$

The negative sign is a reflection of the metric. The other Hamilton's eqns. of motion are

$$\dot{\mathbf{X}}_{\perp} = 2\pi \mathcal{P}_{\perp} \quad \dot{\mathcal{P}}_{\perp} = \frac{\mathbf{X}_{\perp}''}{2\pi} \quad P_+ = \text{const.} \quad (17.39)$$

The required Poisson Brackets are

$$\{X^i, X^j\} = 0 = \{\mathcal{P}^i, \mathcal{P}^j\} \quad \{X^i(\sigma), \mathcal{P}^j(\sigma')\} = \delta^{ij} \delta(\sigma - \sigma') \quad (17.40)$$

In the above, the indices (i, j) are only restricted to the $D - 2$ transverse directions. To facilitate matters further GGRT introduce the mode expansion

$$X^\mu(\sigma, \tau) = q_0^\mu + \sqrt{2}\{a_0^\mu \tau + i \sum_{n \neq 0} \frac{a_n^\mu}{n} \cos n\sigma e^{-in\tau} \quad (17.41)$$

They justify this on the basis of X^μ obeying the equation of motion of Eq. (17.29). The first two terms, the so-called *zero-mode* terms reflect Eq. (17.37). The reality of X^μ imposes $q_0 = q_0^*$, $a_0 = a_0^*$, $a_n^* = a_{-n}$. All the constraints can now be expressed in terms of the normal modes:

$$q_0^+ = 0 \quad a_n^+ = 0 \quad n \neq 0 \quad a_0^+ = \sqrt{2}P_+; \quad (17.42)$$

and,

$$a_n^- = \frac{1}{a_0^+} L_n \quad L_n = \frac{1}{2} \sum_{k=-\infty}^{\infty} a_{-k}^i a_{n+k}^i \quad (17.43)$$

The L_n are the same Virasoro operators we encountered before. Here they arise entirely from the constraints which in turn descended directly from the coordinate conditions. In the summation above, only the transverse modes contribute, which is a big difference from the covariant L_n that appeared in the oscillator description of the dual models. As yet, these are classical considerations, and there are no operators.

In terms of the modes, a_n^i , q_0^i , a_0^+ , q_0^- are the independent variables, all the rest being dependent on these. The Poisson Brackets among these independent variables can be easily worked out from the Poisson Brackets among the independent canonical variables listed already:

$$\{a_n^i, a_m^j\} = -in\delta_{n,-m}\delta^{ij}; \quad \{q_0^i, a_0^j\} = \sqrt{2}\delta^{ij}; \quad \{q_0^-, a_0^-\} = -\sqrt{2}; \quad \{q_0^i, q_0^j\} = 0 \quad (17.44)$$

From these, the PB's of L_n 's can be worked out:

$$\{L_n, L_m\} = -i(n - m)L_{n+m}, \quad \{L_n, a_m^i\} = im a_{m+n}^i, \quad \{q_0^i, L_m\} = \sqrt{2} a_m^i \quad (17.45)$$

It should be noted that the algebra of L_n 's does not have the anomaly or the central charge term; that is a purely quantum effect. GGRT also work out the expressions for the conserved momentum and angular momentum:

$$P^\mu = \frac{1}{\sqrt{2}} a_0^\mu \quad M^{\mu\nu} = \frac{1}{\sqrt{2}} (q_0^\mu a_0^\nu - q_0^\nu a_0^\mu) + i \sum_{n \neq 0} \frac{a_n^\mu a_{-n}^\nu}{n} \quad (17.46)$$

The second of these will play a crucial role in determining the Lorentz invariance of the non-covariant quantization procedure. Finally, the Hamiltonian and the invariant $(\text{mass})^2$ are given by

$$H = 2P_+P_- = L_0 = P^2 + \sum_{n=1}^{\infty} a_n^i a_n^{i*} \quad M^2 = \sum_{n=1}^{\infty} a_n^i a_n^{i*} \quad (17.47)$$

17.3.2 Non-covariant Quantization

Following standard methods, quantization now consists of replacing all Poisson Brackets with commutators; more precisely

$$i\{A, B\} \rightarrow [A, B] \equiv AB - BA \quad (17.48)$$

All the mode-expansion coefficients are now operators, and their commutation relations can be written down:

$$[a_n^i, a_m^j] = n\delta_{n,-m}\delta^{ij}, \quad [q_0^i, a_0^j] = -i\sqrt{2}\delta^{ij}, \quad [q_0^-, a_0^+] = -i\sqrt{2} \quad (17.49)$$

all other commutators among the independent variables vanishing. The a_n for $n > 0$ become annihilation operators and a_{-n} for $n > 0$ become creation operators. This is dictated by the first of the commutation relations above in combination with the requirement of positive-norm states. The fact that operators in general do not commute leads to the well-known *ordering ambiguities* of quantum theories. In the present context, this is most serious for L_0 .

Henceforth, it is understood that where necessary, all expressions are *normal ordered* which means an ordering where all creation operators occur to the left of all annihilation operators. The prescription advocated by GGRT is to replace the occurrence of every unordered L_0 by $L_0 - \alpha_0$ where in the second expression L_0 is normal ordered. This way, the quantum analog of the second constraint becomes

$$a_0^- = \frac{1}{a_0^+} (L_0 - \alpha_0) \quad (17.50)$$

the parameter α_0 will soon be seen to play the role of the universal Regge intercept of the dual models. The invariant $(\text{mass})^2$ is also modified likewise:

$$M^2 = \sum_{n=1}^{\infty} a_n^\dagger \cdot a_n - \alpha_0 \quad (17.51)$$

The algebra of L_n 's undergoes a profound modification:

$$[L_n, L_m] = (n-m)L_{n+m} + \frac{(D-2)}{12} (n^3 - n) \delta_{n,-m} \quad (17.52)$$

The algebra is central extended with the anomaly term! While the (n, m) -dependence of this term is what had been deduced in Chap. 16 only on the basis of Jacobi identity, its coefficient is now seen to be $(D - 2)$ instead of D there. This difference is due to the fact that the present treatment is not Lorentz-covariant and that only the $(D - 2)$ transverse d.o.f contribute, 1 for each d.o.f. Nevertheless, the final results are the same.

Restrictions on α_0 can be seen in many ways. GGRT first analyse the level-1 excitations for which $M_1^2 = 1 - \alpha_0$. Its physical states are transverse. That is compatible with Lorentz covariance only if $M^2 = 0$ which means $\alpha_0 = 1$ which is also the value one gets from the operator formalism of dual models as we saw in Chap. 16.

To summarize, the non-covariant quantization based on light-cone variables admits only the transverse oscillators, and a Fock-space built out of them automatically admits only positive-norm states. Though the level-1 analysis restricts $\alpha_0 = 1$, there seems to be no restriction on D . Of course, $\alpha_0 = 1$ immediately forces the ground state (level-0) to be tachyonic with $M^2 = -1$. Poincare invariance being a feature of the Nambu-Goto action, the theory must be Lorentz-covariant even though it is not *manifestly so*. The way to test this is to work out the Poincare Algebra in terms of the transverse oscillators. The only troublesome case turns out to be $[M^{i-}, M^{j-}]$ which should vanish if the theory is Lorentz-invariant. We display below the final result of GGRT(after correcting a typo):

$$[M^{i-}, M^{j-}] = \frac{1}{4a_0^{+2}} \sum_{m=1}^{\infty} \left\{ m \left(1 - \frac{(D-2)}{24} \right) + \frac{1}{m} \left(\frac{(D-2)}{24} - \alpha_0 \right) \right\} (a_m^i a_{-m}^j - a_m^j a_{-m}^i) \quad (17.53)$$

This can vanish only if $\alpha_0 = 1$, a conclusion the level-1 analysis had already yielded, and $D = 26$. Thus the theory is Lorentz covariant only in 26 dimensions. As already stressed the GGRT analysis vindicate the earlier analyses of Brower [25], Goddard and Thorn [26]; their construction of positive norm Hilbert space coincides with the DDF construction [27] discussed earlier.

17.4 Covariant Quantization

Now we turn to the other method of quantization where Poincare invariance in X^μ is *manifest*. It will turn out to be the same as the operator formalism for the dual models which we have discussed extensively, so a lot of ground work for this approach is already there. In this method, no attempt will be made to explicitly solve the constraints arising out of the coordinate choices and isolate the independent d.o.f.

GGRT go straightaway to the quantum commutation relations without detailed analyses of the classical aspects. However, Goto provides such a classical analysis in [9] a work that preceded that of GGRT by more than a year. We discuss Goto's analyses before going on to quantum theory. Interestingly, Goto does not impose the coordinate conditions from outside but argues for the constraints as arising directly from the invariance properties of the action. This is also the point of view expressed

by Di Vecchia (see the derivations of Eq. (270) of [11]). Nevertheless Di Vecchia does impose the coordinate conditions also to simplify the Euler-Lagrange eqns. to motivate a covariant mode expansion. In contrast Goto does not invoke the coordinate conditions ever, nor make use of the gauge-fixed Euler-Lagrange eqns. (only they have unique solutions, as remarked earlier). All these nuances are very important conceptually. We first discuss Goto's treatment and contrast it with that of GGRT.

Goto makes the pertinent observation that the explicit form of the momentum-currents $\mathcal{P}^\mu$ given in Eq. (17.19) automatically yield (as before we use the notation $\mathcal{P}^\mu$ for $\mathcal{P}_\tau^\mu$):

$$\mathcal{P} \cdot X' \equiv T(\sigma, \tau) = 0 \quad \mathcal{P}^2 + \frac{X'^2}{(2\pi)^2} \equiv H(\sigma) = 0 \quad (17.54)$$

without using any coordinate fixing conditions; they follow just from the definitions of $\mathcal{P}^\mu$ as given by Eq. (17.17). Goto interprets these as *constraints* which are vanishing classically, but need to be interpreted as *weakly vanishing* in quantum theory. From this point onwards Goto turns to the quantum theory. Firstly, he follows Dirac's generalized canonical formalism for treating systems with constraints. Some salient features of Dirac's formalism are (i) determining the *closure* of the algebra of constraints, (ii) checking the consistency of the constraints against time evolution, and, (iii) modifying the naive Hamiltonian by addition of constraints. In the present case, the Naive Hamiltonian vanishes because of the reparametrization invariance of the action, so the Hamiltonian is just made up of constraints.

As far as (i) Goto verifies that

$$[H(\sigma), H(\sigma')] = 8T_0^2 i\delta(\sigma - \sigma') T(\sigma) + 4iT_0^2 \delta(\sigma - \sigma') \frac{\partial T}{\partial \sigma} \quad (17.55)$$

$$[T(\sigma), T(\sigma')] = 2i\delta'(\sigma - \sigma') T(\sigma) + i\delta(\sigma - \sigma') \frac{\partial T}{\partial \sigma} \quad (17.56)$$

$$[T(\sigma), H(\sigma')] = 2i\delta'(\sigma - \sigma') H(\sigma) + i\delta(\sigma - \sigma') \frac{\partial H}{\partial \sigma} \quad (17.57)$$

Goto has taken $[X^\mu(\sigma), \mathcal{P}^\nu(\sigma')] = ig^{\mu\nu} \delta\sigma - \sigma'$ (his metric convention is $g_{00} = 1, g_{ij} = -\delta_{ij}$). Goto specifies a *symmetric ordering* for treating $T(\sigma)$. All these could have been treated classically too with PB's replacing commutators.

The algebra that Goto obtained for $T(\sigma), H(\sigma)$ is the conformal algebra. Nambu, in his [3] also obtains an algebra isomorphic to this. They both lead to the Virasoro algebra among the L_n 's, but with a crucial difference: the absence of the anomaly or the central extension. This is puzzling as the treatments of both Nambu and Goto is quantum mechanical. Goto next introduces what looks like a mode-expansion, but is just a *Fourier Expansion* for $X^\mu, \mathcal{P}^\mu$:

$$X_\mu(\sigma) = \sqrt{\frac{2}{\pi}} \sum_{r=0}^{\infty} \xi_\mu^r \cos r\sigma \quad \mathcal{P}_\mu(\sigma) = \sqrt{\frac{2}{\pi}} \sum_{r=0}^{\infty} \pi_\mu^r \cos r\sigma \quad (17.58)$$

Consequently

$$[\pi_\mu^r, \xi_\nu^s] = i g_{\mu\nu} \delta_{rs} \quad (17.59)$$

It is worth emphasizing that these are just Fourier-Series expansions, and not mode expansions. In fact Goto does not discuss the solutions of the Euler-Lagrange eqns. Goto goes on to introduce the oscillator variables:

$$a_\mu^r = \frac{1}{\sqrt{2r}} (r \xi_\mu^r + i \pi_\mu^r) \quad a_\mu^0 = \frac{1}{\sqrt{2}} \pi_\mu^0 = \frac{P_\mu}{\sqrt{\pi}} \quad (17.60)$$

satisfying

$$[a_\mu^r, a_\nu^{s\dagger}] = -g_{\mu\nu} \delta_{rs} \quad (17.61)$$

To proceed further, Goto introduces the Fourier coefficients of $H(\sigma)$, $T(\sigma)$:

$$H^{(n)} = \frac{1}{2} \int_0^\pi d\sigma \cos n\sigma H(\sigma) \quad T^{(n)} = \int_0^\pi d\sigma \sin n\sigma H(\sigma) \quad (17.62)$$

In terms of $U^{(n)} = H^{(n)} + iT^{(n)}$, $n \geq 1$, Goto obtains the algebra

$$[U^{(n)}, U^{(m)}] = (n - m) U^{(n+m)} \quad (17.63)$$

Though this is similar to the quantum algebra of L_n 's (both are conformal algebras), the crucial anomaly term is missing. The same is true of the algebra obtained by Nambu in [3].

Goto then claims that all physical states must obey $U^{(n)}|phys\rangle = 0$ and $U^{(n)\dagger}|phys\rangle = 0$. Neither Nambu nor Goto draws any D-dependent restrictions.

Returning to GGRT's covariant quantization, the starting point is the covariant commutation relations:

$$[X^\mu(\sigma, \tau), \mathcal{P}^\nu(\sigma', \tau)] = i g^{\mu\nu} \delta(\sigma - \sigma') \quad (17.64)$$

There are several conceptual issues that arise: the “equal time” has been interpreted as “equal τ ”; the meaning of time is somewhat obscure, certainly in comparison to the non-covariant approach. The constraint equations, boundary conditions, the Euler-Lagrange equations and hence the mode expansions are all manifestly covariant. The quantum commutation relations are

$$[a_n^\mu, a_m^\nu] = g^{\mu\nu} \delta_{n, -m} \quad [q_0^\mu, a_0^\nu] = i\sqrt{2} g^{\mu\nu} \quad (17.65)$$

We remind the reader that the GGRT metric convention is $g^{00} = -1$, $g^{ij} = \delta^{ij}$. The Fock-space is built as usual with the oscillators. An immediate problem arises, as has already been discussed at length in the context of the operator formalism for dual models, i.e. an odd number of time-like oscillators can excite negative norm states. This is the price for manifest Lorentz covariance. There, the problem was eventually

resolved through the Virasoro Subsidiary conditions [28]. Nevertheless, the origin of these subsidiary conditions was not totally transparent. But in GGRT and Goto approaches, the Nambu-Goto theory provides these in a natural way as constraints.

The constraints in terms of normal modes become

$$\langle \psi_1 | L_N | \psi_2 \rangle = 0 \quad N \neq 0 \quad \langle \psi_1 | L_0 | \psi_2 \rangle = \alpha_0 \langle \psi_1 | \psi_2 \rangle \quad (17.66)$$

with

$$L_N = \frac{1}{2} \sum_{l=-\infty}^{\infty} : a_{-l} \cdot a_{l+N} : \quad (17.67)$$

the $::$ denote normal ordering. Though this looks a lot like what we encountered in the non-covariant approach, there are differences. In fact, the algebra of the covariant L_N 's is

$$[L_N, L_M] = (N - M)L_{N+M} + \frac{D}{12} \delta_{N,-M}(N^3 - N) \quad (17.68)$$

It has the central extension term but now proportional to D ! This is because all the D -components make equal contributions. As before, the weak vanishing of constraints is equivalent to demanding

$$L_N |phys\rangle = 0 \quad N > 0 \quad L_0 |phys\rangle = \alpha_0 |phys\rangle \quad (17.69)$$

From this point onwards, the problem of ghosts is exactly the one studied by Brower [25], and, Goddard and Thorn [26]. The conclusions are also their conclusions. Let us summarize them here again: For $\alpha_0 = 1$ and $1 \leq D \leq 26$ the theory has no ghosts. When $\alpha_0 < 1$, the restrictions on D are $1 \leq D \leq 25$. However, the DDF construction [27], only in 26 dimensions do they span a complete basis. The $D < 26$ case is handled by the introduction of additional d.o.f over and above those of the classical system to make up exactly 26 d.o.f. These considerations will assume significance when we consider the so called *Effective String Theories* as a description of Yang-Mills flux tubes in arbitrary dimensions.

17.5 The Arvis Quantization

Now we turn to the quantization of Nambu-Goto strings whose end-points are *fixed*. This problem was solved exactly by Arvis in 1983 [29]. This will prove to be extremely relevant to the eventual goal of relating the Static quark-antiquark potential to an effective bosonic string theory.

Apart from the new boundary conditions implied by the fixing of the two ends, the treatment of Arvis closely follows [18]. As we have extensively discussed the latter, it will be straightforward to explain the results of Arvis. He chooses one end

of the string to be at the origin of the X^μ coordinates, while the other end at R^i so the boundary conditions become

$$X^i(0, \tau) = 0 \quad X^i(\pi, \tau) = R^i \quad i = 1 \dots, D-1 \quad (17.70)$$

This is in contrast to the free string whose ends had to move with c while the velocities here are zero! Of the older boundary conditions, coming from the variational principle, the one for X^0 remains the same:

$$X'^0(0, \tau) = X'^0(\pi, \tau) = 0 \quad (17.71)$$

The coordinate conditions and the Euler-Lagrange eqns. remain the same as in the case of the free string:

$$X' \cdot \dot{X} = X'^2 + \dot{X}^2 = 0 \quad \ddot{X}^\mu - X''^\mu = 0 \quad (17.72)$$

However, the new boundary conditions change the mode expansions. The mode expansion for X^0 remains the same as in the GGRT analysis:

$$X^0(\sigma, \tau) = q^0 + P^0 \tau + i\sqrt{2} \sum_{n \neq 0} \frac{a_n^0}{n} \cos n\sigma e^{-in\tau} \quad (17.73)$$

But the mode expansion for X^i changes to:

$$X^i(\sigma, \tau) = \frac{\sigma}{\pi} R^i + \sqrt{2} \sum_{n \neq 0} \frac{a_n^i}{n} \sin n\sigma e^{-in\tau} \quad (17.74)$$

The reality of X^μ imply $a_n^{\mu*} = a_{-n}^\mu$, at the classical level. Introduce the vector a_0^μ :

$$a_0^0 = \frac{P^0}{\sqrt{2}} \quad a_0^i = \frac{R^i}{\sqrt{2}\pi} \quad (17.75)$$

Here and before P^μ is the conserved momentum of the string. The two constraints (coordinate conditions) when expressed in terms of a_n^μ take the forms:

$$a_n \cdot a_0 = -\frac{1}{2} \sum_{n \neq 0} a_n \cdot a_{N-n} \quad a_0 \cdot a_0 + 2 \sum_{n \neq 0} a_n \cdot a_{-n} = 0 \quad (17.76)$$

The paper of Arvis also suffers from typos, confusing notations, and even incomplete eqns! (the first of the constraints is an example). Arvis also opts for a non-covariant approach by fixing the coordinates further with the choices

$$a_n^0 + a_n^1 = 0 \quad q^0 = 0 \quad (17.77)$$

The first of these is identical to what GGRT also choose for their light-cone coordinates. The second is peculiar to the strings with fixed ends. Either way, only transverse degrees of freedom are left as independent variables.

So far, the treatment is classical and the basic Poisson Brackets are

$$\{a_n^i, a_m^j\} = -in \delta_{n,-m} \delta^{ij} \quad (17.78)$$

The second of the constraints will be the determining equation for the energy levels, and is useful to rewrite as

$$(P^0)^2 = \frac{R^2}{\pi^2} + 4L_0 \quad L_0 = \frac{1}{2} \sum_{m \neq 0} a_m^i \cdot a_{-m}^i \quad (17.79)$$

Upon passage to quantum mechanics, the Poisson Brackets become commutators

$$[a_n^i, a_m^j] = n \delta_{n,-m} \delta^{ij} \quad (17.80)$$

With this, the same ordering ambiguities as arose in the GGRT analysis arise here too and the remedies are same: to treat all operators as normal ordered and allow for a shift in L_0 so that Eq. (17.79) is modified to:

$$(P^0)^2 = \frac{R^2}{\pi^2} + 4(L_0 - \alpha_0) \quad L_0 = \sum_{m \neq 0} a_m^{i\dagger} \cdot a_m^i \quad (17.81)$$

As in the GGRT analysis, Arvis fixes α_0 by requiring Rotational invariance, i.e. by requiring the correct algebra of the rotation generators. However, additional complications arise due to the presence of the 'external' vector R^i . Though these are not dynamical variables, Arvis introduces variables S^j conjugate to them in the proof. This does not seem entirely convincing and perhaps a more refined analysis is in order. The situation is entirely analogous to, say, an electron in an external magnetic field. If the magnetic field is held fixed, there is obviously a violation of rotational invariance of the electronic system. The correct way to discuss the overall rotational invariance is to change the external magnetic field along with rotating the electron coordinates and momenta. Arvis's method is a formal way of realizing this. Accepting his arguments, rotational invariance fixes $\alpha_0 = \frac{(D-2)}{24}$, $D = 26$, exactly as in the GGRT case. Putting everything together one gets for the ground state energy

$$V(R) = \left\{ \frac{R^2}{4\pi^2 \alpha'^2} - \frac{(D-2)}{24\alpha'} \right\}^{\frac{1}{2}} \quad (17.82)$$

The excited state energies can likewise be worked out. Though formally it looks as if this expression is for any D , rotational invariance restricts its use to only $D = 26$. This point will occupy us greatly later on. The Arvis potential turns complex for $R < R_c = \pi \sqrt{\frac{(D-2)\alpha'}{6}}$. This has the same physical meaning as the tachyonic nature of the free string ground state.

17.6 Path Integral Quantizations

As already stressed, the Nambu-Goto action is a world-sheet action that is invariant under general transformations of the world-sheet coordinates. What distinguishes the general coordinate invariance in this case from invariance under general coordinate transformations in General Relativity or Riemannian geometry is that the action is invariant even without an *intrinsic* world-sheet metric. However, there is a formulation of string actions, due to A.M. Polyakov, which explicitly includes an intrinsic metric for the world sheet, with the action:

$$S(X, h) = \int d^2\xi \sqrt{-h} h^{\alpha\beta} \partial_\alpha X \cdot \partial_\beta X \quad (17.83)$$

where $h^{\alpha\beta}$ is the intrinsic metric on the two-dimensional world-sheet. This action is also invariant under general transformations of the world-sheet coordinates, and this time the invariance is of exactly the same type as in General Relativity, and, Riemannian Geometry. As such this invariance is devoid of physical significance in the sense that any theory can be made invariant this way. Another way to see this is that this invariance has no effect on the physical degrees of freedom.

At a classical level, the intrinsic metric can be eliminated in favour of the induced metric as a result of the h -eqns of motion and one recovers the Nambu-Goto action. This type of action was first proposed by Brink, Di Vecchia and Howe [30], and, by Deser and Zumino [31]. However, a very influential work by Polyakov [32] based on such actions has revealed the quantum aspects. In particular, that only in $D = 26$ is the approach equivalent to dual models. More importantly, Polyakov's approach shed a lot of light in $D < 26$. Polyakov based his work on the path-integral approach. The above-mentioned action is more popularly called the *Polyakov Action*. A Hamiltonian approach was investigated by Date, Sumitra and the author [33].

In addition to the general coordinate invariance the Polyakov action is invariant under *local* Weyl scalings of h of the form $h^{\alpha\beta} \rightarrow \lambda(\xi) h^{\alpha\beta}$ (X^μ are not affected by these). In fact these are the invariances with non-trivial physical content. Of course, the Poincare transformations among X^μ are also the invariances of the theory. An elaborate treatment of the path-integral quantization based on the Polyakov action can be found in Polchinski's book *String Theory, Vol I* [34]. We shall not go further into these details here, but mention that the Polykov formulation will play a very important role in our discussions of Effective string theories later on.

References

1. H. Poincare, *Science and Hypothesis* (1902)
2. Y. Nambu, in *Proceedings of International Conference on Symmetries and Quark Models*, ed. by R. Chand (Wayne State University, Gordon and Breach, 1970)
3. Y. Nambu, *Duality and Hadrodynamics*. Lectures at the Copenhagen Symposium, 1970, reprinted in *Broken Symmetry: Selected Papers of Y. Nambu*, ed. by T. Eguchi, K. Nishijima (World Scientific, 1995)

4. L. Susskind, Phys. Rev. Lett. **23**, 545 (1969)
5. L. Susskind, Phys. Rev. D. **1**, 1182 (1970)
6. L. Susskind, Nuovo Cim. **69**, 210 (1970)
7. H.B. Nielsen, Paper submitted to the 15th Int. Conf. on High Energy Physics, Kiev, 1970; Nordita Preprint (1969)
8. H.B. Nielsen, String from Veneziano model, [arXiv:0904.4221v1](https://arxiv.org/abs/0904.4221v1) [hep-ph]
9. T. Goto, Prog. Theor. Phys. **46**, 1560 (1971)
10. P.H. Frampton, *Dual Resonance Models* (W.A. Benjamin Inc, 1974)
11. P. Di Vecchia, The birth of string theory, [arXiv:0704.0101v1](https://arxiv.org/abs/0704.0101v1) [hep-th]
12. N.D. Hari Dass, Lattice Theory for nonspecialists (1984), NIKHEF Preprint NIKHEF-H-84-11
13. L.D. Landau, E.M. Lifshitz, *Theory of Elasticity (Course of Theoretical Physics)*, vol. 7. (Elsevier Publishing Company, 1986)
14. E. Kretschmann, Ann. Phys. Leipzig. **53**, 575 (1917)
15. N.D. Hari Dass, P. Matlock, (2007), [arXiv:0709.1765](https://arxiv.org/abs/0709.1765) [hep-th]
16. N.D. Hari Dass, P. Matlock, Indian J. Phys. **88**, 965 (2014)
17. J. Polchinski, A. Strominger, Phys. Rev. Lett. **67**, 1681 (1991)
18. P. Goddard, J. Goldstone, C. Rebbi, C. Thorn, Nucl. Phys. **B56**, 109 (1973)
19. G. Barton, Introduction to advanced field theory, in *Interscience Tracts on Physics and Astronomy*, vol. 22 (1963)
20. S. Weinberg, *The Quantum Theory of Fields-I* (Cambridge University Press)
21. J.D. Bjorken, S. Drell, *Relativistic Quantum Fields* (McGraw Hill Publishers)
22. C. Itzykson, J.-B. Zuber, *Quantum Field Theory* (McGraw Hill Publishers)
23. J. Polchinski, TASI lectures on D-branes, in *Fields, Strings and Duality, TASI 1996* (World Scientific Publishers)
24. P.A.M. Dirac, Can. J. Math. **2**, 129 (1950); Proc. R. Soc. A **246**, 326 (1958)
25. R.C. Brower, Phys. Rev. D. **6**, 1655 (1972)
26. P. Goddard, C.B. Thorn, Phys. Lett. **B40**, 235 (1972)
27. E. Del Giudice, P. Di Vecchia, S. Fubini, Ann. Phys. **70**, 378 (1972)
28. M.A. Virasoro, Phys. Rev. D. **1**, 2933 (1970)
29. J.F. Arvis, Phys. Lett. B. **127**, 106 (1983)
30. L. Brink, P. Di Vecchia, P. Howe, Phys. Lett. B. **65**, 471 (1976)
31. S. Deser, B. Zumino, Phys. Lett. B. **65**, 369 (1976)
32. A.M. Polyakov, Phys. Lett. B. **103**, 207 (1981)
33. N.D. Hari Dass, Hamiltonian formulation of Polyakov gravity, in *Modern Quantum Field Theory*, ed. by S. Das et al., (World Scientific, 1991)
34. J. Polchinski, *String Theory*, vol. 1. Cambridge University Press

Part III

Strings Lost: QCD, The Field Theory of Strong Interactions



18.1 Introduction

In this chapter we explore the concept of *Effective Field Theories* (EFT). The motivations are twofold: one to show how impressive progress was made in our understanding of strong interactions well before *Quantum Chromodynamics* (QCD) got created and accepted as a relativistic quantum field theory of strong interactions. The other has to do with the main theme of this book, namely, an *effective string* description of Yang–Mills flux tubes. This chapter should enable the reader to appreciate the nuances of such a description.

Effective descriptions have been part of our scientific world-view all along. Classic examples are elasticity and fluid dynamics [1, 2]. The modern basis for these phenomena lies in electromagnetism, including its quantum aspects. But long before the modern ideas crystallized into successful theories (read Quantum Mechanics and Quantum Electrodynamics), both elasticity and fluid mechanics could account for an impressive range of phenomena. This illustrates the spirit and essence of effective descriptions. They typically consist of some basic concepts and a few parameters whose ultimate explanations come from deeper fundamental theories. In the case of elasticity, this effective description is in terms of notions like stress and strain tensors, Hooke's law and a free energy whose stationarity explains myriad elastic phenomena. Likewise, the Navier–Stokes description of fluid behaviour. They are, at a more microscopic level, consequences of inter-atomic forces like Van der Waals attraction, etc. Another example of effective descriptions is from optics [3]. With effective parameters such as refractive index and absorption coefficient, possibly frequency dependent, myriads of optical phenomena can be understood. At the microscopic level they need theories of scattering, resonance, etc. Curiously, as explained in Chap. 4 of this book, it is a description of these that led Kramers and Kronig to discover the connection between causality and analyticity.

Eventually, we wish to stress the role played by *symmetries* in the construction of effective descriptions. This aspect is not particularly evident in the examples we have highlighted so far. Another aspect of effective descriptions that is very important is their range of validity. In all the examples above, clearly the descriptions are not valid at atomic scales. In the next two sections we take up two examples, namely, the Fermi theory of weak interactions, and, Superconductivity that will exhibit the important roles of symmetry. In superconductivity at a much deeper level than in the Fermi theory and its subsequent improvements. Finally, in the last section, we take a careful look at the effective field descriptions of strong interactions at low energies.

18.2 Effective Description of Weak Interactions

A breakthrough in the description of beta-radioactivity was Enrico Fermi's theory of beta-decay [4] which boldly incorporated Pauli's neutrino idea. The microscopic description of this phenomenon is currently given by the *Standard Model* which took close to four decades for its development from the time of Fermi's theory. Nevertheless, Fermi's theory and its many important generalizations and improvements that followed accounted for indeed a very large class of phenomena. It is in this sense that these descriptions are *effective* descriptions. We give here a brief account of these with particular emphasis on their effective field theory aspects.

The original Fermi theory described beta-decay of neutrons $n \rightarrow p + e + \bar{\nu}$ by the effective Hamiltonian

$$H_W = C_V \int d^3x \bar{\psi}_p(x) \gamma^\mu \psi_n(x) \cdot \bar{\psi}_e(x) \gamma_\mu \psi_\nu(x) \quad (18.1)$$

with C_V a measure of the strength of interactions (to be related to the Fermi-coupling G_F shortly), with mass dimensions $[M]^{-2}$. This was to be experimentally determined by, say, the observed neutron lifetime (whose value, incidentally, has undergone big changes! Even currently there is a discrepancy of nearly 8s out of 880s between the so-called *beam* and *bottled* neutrons). In proposing this, Fermi had evoked an analogy with electromagnetism where interactions, albeit of long-range, can be described in terms of a current-current structure. The symmetries Fermi had originally assumed were Lorentz invariance and Parity invariance. It was soon realized on the one hand that Fermi theory could not accommodate nuclear beta-decays where the spin of the nucleus changed, and on the other hand that Lorentz Invariance along with Parity and Time-reversal invariances (believed to be good in those days) actually allowed for a generalization of the Fermi Hamiltonian to include Vector, Axial vector, Scalar, Pseudoscalar, as well as Tensor interactions with respective strengths of C_V, C_A, C_S, C_P, C_T .

The experimental status of these remained confusing and inconclusive for a long time. We refer the reader to Chaps. 29–31 of [5] for a critical analysis. An analysis by Ruderman and Finkelstein [6], as early as 1949, of charged pion decays of the type $\pi^\pm \rightarrow l^\pm + \nu_l$, where l refers to the leptons, i.e. e, μ , gave strong indications that

C_P was highly suppressed, a consequence of which was that the decay into the lighter electrons is several orders of magnitude weaker than the one to muons. The phase space volumes are more or less the same in the two cases! The perturbative treatments of this process were plagued by divergences. This period of confusion ended with two dramatic developments: first was the startling and revolutionary discovery by Yang and Lee [7] that there existed no compelling evidence for Parity invariance, and the subsequent experimental discovery by C. S. Wu that in fact in beta-decay, parity was maximally violated [8]. Even with this mind-boggling development, the nature of beta-interactions was far from clear. Another revolutionary development completely cleared the air, and that was the discovery of $V - A$ structure of weak currents by Sudarshan and Marshak [9], by Feynman and Gell-Mann [10], and by Sakurai [11]. With this, Fermi's original Hamiltonian density for neutron beta-decay was changed to

$$\mathcal{H}_\beta = \frac{G_V}{\sqrt{2}} \bar{\psi}_p \gamma^\mu \left(1 - \frac{G_A}{G_V} \gamma_5\right) \psi_n \cdot \bar{\psi}_e \gamma_\mu (1 - \gamma_5) \psi_\nu \quad (18.2)$$

While the leptonic part is precisely of the $V - A$ type, the hadronic part has a $\frac{G_A}{G_V}$ part. This will be of great significance as we go along. Another idea that gained a lot of traction was that weak interactions, like gravitation, are also *Universal*. This prompted that even $\mu \rightarrow e + \nu + \bar{\nu}$ should also be described by a Hamiltonian density of the form (after due recognition of the $V - A$ form for leptons)

$$\mathcal{H}_\mu = \frac{G_\mu}{\sqrt{2}} \bar{\psi}_\nu \gamma^\mu (1 - \gamma_5) \psi_\mu \cdot \bar{\psi}_e \gamma_\mu (1 - \gamma_5) \psi_\nu \quad (18.3)$$

Though two couplings appear, experimentally it was found that $G_V = G_\mu$ to about a per cent! Even that very small deviation is rooted in the so-called *Cabibbo Angle* whose value is $\theta_c = 0.26$ such that $G_V = G_\mu \cos \theta_c$ [12]. Feynman and Gell-Mann likened the near equality of G_V and G_μ to the equality of electrical charges of, say, proton and electron, despite the fact that protons can experience strong interactions while electron cannot. Just as the equality of electric charges of all particles is a consequence of the conservation of the electric current, the equality of G 's, according to them, was a consequence of the conservation of the isospin vector current. This is the famous *Conserved Vector Current* (CVC) hypothesis. This will play a crucial role for our discussions of the last section of this chapter. Experimentally, in the early days it had been observed that $\frac{G_A}{G_V} \approx 1.18$ [13] based on the then prevailing value of 11.7 min for the lifetime of neutrons. The significance of this value of $g_A \equiv \frac{G_A}{G_V}$ will occupy us greatly in the last section.

The “current-current” picture was soon extended to the decays of strange particles also. It is in this context that the Cabibbo angle assumes its significance; while the strangeness conserving currents enter with strength $\cos \theta_c$, the strangeness changing currents enter with strength $\sin \theta_c$ making them rather rare. There are many selection

rules governing the strangeness changing processes which we shall not go into except to state that they have natural explanations in the *Quark Models* to be discussed in the next chapter. The reader is referred to the book by Gasiorowicz [5] for an excellent coverage.

The upshot of all this is that the beta-decay Hamiltonian density is still of the current-current form (the overall sign depending on the metric convention):

$$\mathcal{H}_w = \frac{G_F}{\sqrt{2}} J^\mu(x) J_\mu^\dagger(x) \quad (18.4)$$

with G_F being the *Fermi Coupling Constant* (numerical value $1.026 \times 10^{-5} m_p^{-2}$), and, where $J_\mu = l_\mu + h_\mu$ with first term for leptonic currents and the second for hadronic currents. Furthermore,

$$\begin{aligned} l_\mu(x) &= \bar{\psi}_e(x) \gamma_\mu (1 - \gamma_5) \psi_{\nu_e}(x) + \bar{\psi}_\mu(x) \gamma_\mu (1 - \gamma_5) \psi_{\nu_\mu}(x) \dots \\ h_\mu(x) &= \cos \theta_c h_\mu^{\Delta S=0} + \sin \theta_c h_\mu^{\Delta S=1} \end{aligned} \quad (18.5)$$

with θ_c being the Cabibbo angle. The hadronic currents also display a $V - A$ structure:

$$h_\mu^{\Delta S=0} = V_\mu^{\Delta S=0} - A_\mu^{\Delta S=0} \quad h_\mu^{\Delta S=1} = V_\mu^{\Delta S=1} - A_\mu^{\Delta S=1} \quad (18.6)$$

Sometimes, the hadronic currents are explicitly written out in terms of hadronic field operators, as in

$$h_\mu^{\Delta S=0} = \bar{\psi}_p(x) \gamma_\mu (1 - \gamma_5) \psi_n(x) + \dots \quad (18.7)$$

with the dots denoting contributions from other hadrons than the nucleon doublet. This is really not necessary as in the end the matrix elements of these currents between hadronic states are inferred on general principles of Lorentz Invariance, conservation or partial conservation, quantum numbers, etc. To illustrate this, and also the apparent puzzling feature that in the hadronic current the factor $g_A = \frac{G_A}{G_V}$ does not explicitly occur, let us write down the matrix elements of the strangeness preserving hadronic current between nucleon states

$$G_F \langle p | h_\mu^{\Delta S=0} | n \rangle = G_V \bar{u}_p \gamma_\mu \left(1 - \frac{G_A}{G_V} \right) u_n \quad (18.8)$$

which makes contact with the “phenomenological” form of Eq. (18.2).

The generalized current-current weak interactions of Eq. (18.4) provide very good phenomenological descriptions of a whole variety of weak interaction phenomena such as (i) purely leptonic processes like μ -decay, (ii) semi-leptonic processes like beta-decay, (iii) non-leptonic decays like $K^0 \rightarrow \pi^+ + \pi^-$ (the reader is referred to [5] for details) and (iv) processes like muon capture by nuclei $\mu^- + Z \rightarrow (Z - 1) + \nu$, and many more! This effective description has clearly relied on various aspects of symmetries, or their lack of, that govern them. Finally, we address the

question of the range of validity of the Fermi effective descriptions. This is best exemplified by the high energy behaviour of, say, electron-neutrino or neutrino-nucleon scatterings. The scattering amplitudes given by the effective theories grow too fast, linearly with energy, something incompatible with unitarity. Therefore at energies comparable to $\sqrt{G_F^{-1}} \approx 250 \text{ GeV}$, the effective descriptions are no longer effective.

The microscopic theory of all these processes, namely, the electro-weak unification theories, pioneered by S. Glashow, A. Salam, J. C. Ward, S. Weinberg, G. 't Hooft and M. Veltman, was fully formulated only some 40 years after Fermi theory (for a thorough discussion see Weinberg's book [14] Sect. 21.3). That should put in proper perspective this discussion about effective descriptions. The standard model has much of the same symmetry content as the effective descriptions above (more or less by construction). It has, however, entirely new aspects such as *Spontaneous Breaking of Global Symmetries* and consequences of gauging such systems. This aspect will also be shared by Superconductivity, as will be explained in the next section. This parallel between elementary particles and superconductivity was the basis for path-breaking developments by Nambu and his collaborators (more on this in the last section).

In terms of the current-current picture, the microscopic standard model sprang a big surprise! According to it, the weak interactions are also mediated by *Neutral Currents* (the currents considered so far are *Charged Currents*; for example, in beta-decay, the nucleon charge changes by one unit, i.e. $\Delta Q = 1$). According to Weinberg (see Ref. 9 of the Chap. 21 of [14]), such neutral currents had been speculated well before the birth of the Standard Model by Teller, Kemmer, Wentzel, Bludman, Leites-Lopes and perhaps more. One of the reasons for their not appearing in the effective descriptions had to do with the extreme difficulty of seeing their effects in hadronic decays. Past their prediction, they were first seen in $\bar{\nu}_\mu - e^-$ scattering in 1973 [15].

18.3 Effective Descriptions of Superconductivity

Superconductivity was serendipitously discovered by Kammerlingh Onnes and his students somewhere around 1911. It is the phenomenon by which metals completely lose their electrical resistance below a certain transition temperature T_c . Another striking feature, discovered in 1933 by Meissner and Ochsenfeld [16], was that superconductors completely expelled magnetic flux. This was instrumental in the proposal of the Dual Superconductor mechanism for Quark Confinement, discussed in Chap. 19. The microscopic theory was given by Bardeen, Cooper and Schrieffer [17]. Superconductivity is discussed in several excellent resources because of its extreme importance both for technology and science, so we shall not undertake that exercise here. There is also an excellent exposition by Weinberg in his book [14] (Sect. 21.6). Our purpose is to highlight how effective descriptions provided deep explanations long before the microscopic BCS theory came out, and to emphasize the symmetry and symmetry breaking aspects of it. According to Weinberg, *Super-*

conductivity is an exceptionally enlightening example of the power of effective field theories. (Opening paras of Sect. 21.6 of [14]). The effective description we have in mind is the *Ginzburg–Landau Theory* that was formulated in 1950 [18]. An extensive review of Ginzburg–Landau description can be found in [19].

As succinctly stated by Weinberg, most condensed matter textbooks focus more on the detailed dynamical aspects of superconductivity, rather than emphasizing the broken symmetry aspects of it. Weinberg’s beautiful one line description of superconductivity is *A superconductor is simply a material in which electromagnetic gauge-invariance is spontaneously broken.* Weinberg has elaborated on the symmetry breaking aspects in [20]. However, Nambu [21], and, Anderson [22] did emphasize these important aspects. Unfortunately, despite the wide use of the phrase *spontaneous symmetry breaking of gauge symmetry* even by leading experts, gauge-invariance is not a *symmetry*, and there can be no such thing as it’s spontaneous breaking. Instead what most people have in mind is a spontaneous breaking of global gauge-invariance, which is indeed a symmetry, and the consequences of gauging such a system. There is enough to write a book on this, but we shall not elaborate further here, except for highlighting them as we go along.

Let us start with the Ginzburg–Landau (GL) discussion in the absence of electromagnetism first. They introduced an *Effective Field*, called *Order Parameter* by condensed matter physicists, which is a complex valued $\psi(x)$. To begin with, the system is endowed with the *global* invariance

$$\psi(x) \rightarrow \psi'(x) = e^{iq^* \lambda(x)} \psi(x) \quad (18.9)$$

Here q^* has the interpretation of an effective charge. One then introduces a globally gauge-invariant *Free Energy density* (the equivalent of a Hamiltonian density, motivated more by thermodynamic rather than by time-evolution context)

$$\mathcal{F} = \frac{\alpha(T)}{2} |\psi|^2 + \frac{\beta(T)}{4} |\psi|^4 + \gamma(T) |\nabla \psi|^2 \quad (18.10)$$

The last term can also be written in the more suggestive form $\gamma(T) = \frac{1}{4m^*(T)}$ with $m^*(T)$ having the interpretation of a temperature dependent effective mass (in what follows the units are chosen such that $\hbar = c = 1$). The reason for choosing 4 in the denominator in place of the customary 2 will be clear shortly. The Free energy itself is the volume integral of $\mathcal{F}$:

$$F(\psi) = \int d\mathbf{r} \mathcal{F} \quad (18.11)$$

and is a *functional* of the order parameter.

At this point, we add a few words about global gauge-invariance being a *symmetry* in the sense of Wigner. Under a global gauge transformation, one gets two *physically distinct* order parameter configurations ψ, ψ' . This is completely analogous to rotations; under rotations too one obtains two distinct configurations or “states”, but if there is rotational invariance, their energies are the same, or more precisely, states

connected by rotations are *degenerate*. Likewise, the distinct order parameter configurations related by global gauge transformations are *degenerate* because $\mathcal{F}$ is global gauge-invariant. It is the degeneracy of states that is the hallmark of a symmetry.

The meaning of the order parameter here is that it vanishes in the normal phase and takes on non-zero values in the superconducting phase. In order to understand the nature of the phase transition, it suffices to study the GL theory in the vicinity of T_c . We shall not go into the technical complications that extremely close to T_c GL type descriptions (also called *Mean Field Theories*) break down. Now, as usual, equilibrium conditions are sought by minimizing the free energy F , leading to the familiar Euler-Lagrange variational methods. For homogeneous solutions, $\gamma(T)$ is not relevant. In order to achieve zero values of the order parameter for $T > T_c$, and non-zero values below T_c , GL made the simple ansatz that near T_c , $\alpha(T) = a(T - T_c)$ with a positive and, $\beta(T) = b$ with b also positive constant. Then for $T > T_c$, F is minimized for $|\psi|^2 = 0$, while for $T < T_c$, $|\psi|^2 = \frac{a}{b} \cdot (T_c - T)$. This leads to a jump in free energy density from zero to $\frac{F}{V} = -\frac{a^2(T - T_c)^2}{4b}$, signifying a *Second Order Phase Transition*.

In the normal phase, also called the *Symmetric Phase*, characterized by $|\psi| = 0$, the state of minimum F is unique and the phase is meaningless. But in the superconducting phase, characterized by $|\psi| \neq 0$, the phase of ψ becomes meaningful. The uniqueness of the ground state now requires a definite value of the phase. But ground states with different phases all have the same F . Thus the ground state, also called the vacuum state, is *infinitely degenerate*. Consequently, the ground state is no longer invariant under global gauge transformations. Nevertheless, the Euler-Lagrange equations are all *covariant* under the symmetry transformations, as can be easily verified. Such a phenomenon is called *Spontaneous Symmetry Breaking* (SSB).

The $|\nabla \psi(x)|^2$ term in $\mathcal{F}$ gives rise to a term of the form $|\psi|^2 (\nabla \phi(x) \cdot \nabla \phi(x))$. This shows that long wavelength fluctuations in this phase can have arbitrarily low energies. Particle physicists would think of them as *zero mass* excitations. This is the essential content of the so-called Goldstone modes (particles) as remarked by J. Goldstone for the first time [23]. This was proved as a theorem by Goldstone, Salam and Weinberg [24]. The relevance of these concepts was highlighted by Anderson [22] and Nambu [21]. Heisenberg was the first to realize the importance and consequences of SSB. He even tried to interpret pions as Goldstone modes of SSB, unfortunately of the wrong symmetry, namely, Isotopic Spin. That would have led to scalar pions which are in fact pseudo-scalars. The correct picture was given by Nambu as will be explained in the last section. The same sort of symmetry breaking happens in Ferromagnets also (where Heisenberg first encountered it).

Considering situations with spatial variations, $\gamma(T)$ is assumed to be constant γ ; this then leads to the expression for the *correlation length* ξ to be $\xi^2 = \frac{\gamma(T)}{\alpha(T)} = \frac{\gamma}{a} (T_c - T)$. In other words, the correlation length diverges as $(T_c - T)^{-1/2}$ with the characteristic mean field exponent of $\frac{1}{2}$.

Now we briefly discuss the behaviour of superconductors in a magnetic field. This amounts to coupling the complex order parameter field to electromagnetic fields,

also called *gauging*, as the global gauge-invariance is now promoted to *local gauge-invariance*. Only magnetic fields are considered. The expression for $\mathcal{F}$ is obtained by the standard minimal substitution:

$$\mathcal{F}_{em} = \alpha(T) |\psi(x)|^2 + \beta(T) |\psi(x)|^4 + \frac{1}{4m^*(T)} |(\nabla + i q^* \mathbf{A})\psi|^2 + \frac{B^2}{8\pi} \quad (18.12)$$

with $\mathbf{A}$ being the vector potential and the magnetic field given by $\mathbf{B} = \nabla \times \mathbf{A}$.

The gauge-invariance now is *local* in the sense that $\mathcal{F}$ is invariant under

$$\psi(x) \rightarrow \psi'(x) = e^{i q^* \lambda(x)} \psi(x) \quad \mathbf{A}(x) \rightarrow \mathbf{A}'(x) = \mathbf{A}(x) - \nabla \lambda(x) \quad (18.13)$$

Under this $\mathbf{B}(x)$ remains invariant. Now the nature of gauge-invariance has completely changed. $\psi(x)$ and $\psi'(x)$ are no longer *distinct* physical configuration but different descriptions of the *same* physical configurations. This is better appreciated by looking at the electromagnetic fields $\mathbf{E}(x)$, $\mathbf{B}(x)$. The vector potentials $\mathbf{A}(x)$, $\mathbf{A}'(x)$ represent the *same* physical electromagnetic fields. So the local gauge-invariance of $\mathcal{F}$ does not represent any degeneracy anymore and consequently is no longer a symmetry in the Wignerian sense, and the phrase *Gauge Symmetry* is really an *oxymoron*!

What is more, *Spontaneous Breaking of Gauge Symmetry* is a double oxymoron, as there was no symmetry to break to begin with. In the superconducting phase with $|\psi| \neq 0$, which is now a locally gauge-invariant statement, the phase $\phi(x)$ no longer has any physical meaning as it can be transformed away *everywhere* through local gauge transformations, unlike the case of the global gauge invariance where it can only be transformed away at one point (which can however be chosen at will). Equivalently, configurations with different phases are to be identified with each other now, and there is no longer any degeneracy of states. Yet another way to see this is that Noether's theorem does not associate any *conserved quantities* with local gauge invariances.

Minimization of F w.r.t $\mathbf{A}(x)$ yields:

$$\nabla \times \mathbf{B}(x) = 4\pi \mathbf{j} \quad \mathbf{j}(x) = -\frac{q^*}{2m^*} |\psi|^2 (\nabla \phi(x) + q^* \mathbf{A}(x)) \quad (18.14)$$

So far we left the value of q^* unspecified. On phenomenological grounds the preferred value for $q^* = 2e$. This gets further support from the *Cooper Pair* mechanism as the origin of BCS-type superconductors. According to Leon Cooper, pairs of electrons condense and this is the dynamical origin of superconductivity [25]. This also motivates the choice of $\gamma(T) = \frac{1}{4m^*(T)}$ instead of the expected $\frac{1}{2m^*(T)}$. The second of the Eq. (18.14) is the essential content of the famous *London Equation* introduced in 1933 by the London brothers Fritz and Hainz London [26] in a phenomenological attempt to explain the Meissner effect. Their starting gauge-invariant equations were

$$\nabla \times \mathbf{j} = -\frac{e^2 n_s}{m} \mathbf{B} \quad \frac{\partial \mathbf{j}}{\partial t} = \frac{e^2 n_s}{m} \mathbf{E} \quad (18.15)$$

where n_s is the density of the superconducting phase. These were eventually combined into one equation, in the so-called London gauge $\nabla \cdot \mathbf{A} = 0$ into

$$\mathbf{j}(x) = -\frac{e^2 n_s}{m} \mathbf{A}(x) \quad (18.16)$$

The gauge invariance aspects of the London equation have led to the massive amount of confusion! The first, naive, thinking that this equation is not gauge invariant because of the explicit appearance of $\mathbf{A}(x)$ is however obviously incorrect as it is to hold only in the London gauge. Even that hides the true meaning of gauge-invariance in a superconductor. Before that, let us see how the London equation accounts for the Meissner–Ochsenfeld effect. On combining with Ampere’s law $\nabla \times \mathbf{B} = \mathbf{j}$ one gets

$$\nabla^2 \mathbf{B}(x) = \frac{4\pi n_s e^2}{m} \mathbf{B}(x) \quad (18.17)$$

The physically meaningful solutions of this are exponentially decaying

$$\mathbf{B}(x) = e^{-\frac{x}{\lambda}} \mathbf{B}(0) \quad \lambda^{-2} = \frac{4\pi n_s e^2}{m} \quad (18.18)$$

displaying the *Penetration Depth* and the attendant Meissner effect. Now let us come to the second of Eq. (18.14):

$$\mathbf{j}(x) = -\frac{q^*}{2m^*} |\psi|^2 (\nabla \phi(x) + q^* \mathbf{A}(x)) \quad (18.19)$$

On ignoring the $\nabla \phi$ term and identifying $|\psi|^2$ with $\frac{n_s}{2}$, $q^* = 2e$, one recovers the London equation including the correct sign. But this equation was obtained *without* making any gauge choices, while the London equation was to be valid only in the London gauge! Also, is there any justification for neglecting the $\nabla \phi(x)$ term? In arriving at Eq. (18.19) no assumptions of homogeneity were made. On the other hand, we have already pointed out that due to local gauge-invariance, the phase $\phi(x)$ has no invariant meaning and it is possible to find a gauge where $\phi(x)$ (not it’s gradient) can be made to vanish everywhere. In such a gauge, Eq. (18.14) indeed becomes the London equation exactly. This gauge can be called the *Ginzburg–Landau–London* (GLL) gauge. While the London gauge is defined through conditions on $\mathbf{A}$, GLL-gauge is defined through conditions on the phase of the order parameter.

But more importantly, there is no need to fix any gauge at all. Because of Eq. (18.13), the combination $(\nabla \phi(x) + 2e \mathbf{A})$ is manifestly gauge-invariant and Eq. (18.19) is the manifestly gauge-invariant form of the London equation.

This observation leads to a radical interpretation of superconductivity in the presence of electromagnetic fields! To see that, let us return to Eq. (18.12) and recognize that the $\gamma(T)$ term can be rewritten as

$$\gamma(T) |\psi(x)|^2 (\nabla \phi(x) + 2e \mathbf{A}(x))^2 \quad (18.20)$$

This suggests introducing the gauge-singlet field $\mathbf{C}(x) \equiv \mathbf{A}(x) + \frac{1}{2e} \nabla \phi(x)$, in terms of which $\mathbf{B}(x) = \nabla \times \mathbf{C}(x)$. Amazingly Eq. (18.12) can be rewritten entirely in terms of the gauge-singlet fields $|\psi|$, $\mathbf{C}(x)$ (modulo terms depending on $\nabla|\psi|$):

$$\mathcal{F}_{em} = \alpha(T)|\psi(x)|^2 + \frac{\beta(T)}{2} |\psi(x)|^4 + \frac{e^2 |\psi(x)|^2}{2m} C^2 + \frac{(\nabla \times \mathbf{C})^2}{8\pi} \quad (18.21)$$

This represents a theory of *massive* vector fields in the superconducting phase, with the mass of the vector-field equalling the inverse of the penetration depth! Gauge invariance has completely disappeared!

There is still an order parameter $|\psi|$ that distinguishes between normal and superconducting phases, but it is *real*! In the passage from normal phase where it is zero to the superconducting phase where it is non-zero, there is no spontaneous symmetry breaking of any kind. In the $\mathbf{A}$, ψ description there were two degrees of freedom each from them; in the $\mathbf{C}$, $|\psi|$ description, the massive vector field contributes 3 and $|\psi|$, 1, so total is still 4.

This is nothing but the Anderson–Higgs effect, also discovered by Guralnik, Hagen, Kibble, and perhaps by more. This is the correct gauge invariant description of superconductivity.

18.4 Effective Descriptions of Strong Interactions

In this last section, we shall look at the deep and fascinating development of effective descriptions of strong interactions long before the inception of Quantum Chromodynamics (QCD). In fact, these developments were the very bedrock for the successful construction of QCD. The effective field theory of strong interactions is a vast area with continuous developments for the past seven decades, including the present. So, it is beyond the scope of this section to do adequate justice to it. What I hope to do is give a succinct but conceptually and technically accurate account. For most part, I will follow my own lecture notes [27]. See also the chapter on effective field theories in the book by Weinberg [14] (Sects. 19.5–19.8) for a lucid and extensive coverage of all important aspects. Another great source is Nambu’s article [28], reprinted in [29]. Both of them come with numerous references to aid the reader further.

After the confidence in meson field theories ebbed in the early 1950s, numerous alternatives to an understanding of strong interactions were pursued. Among them, the S-matrix approach and Dispersion Relations have been extensively discussed in the earlier chapters of the book. Low [30], and Gell-Mann and Goldberger [31] pioneered the techniques of *Low Energy Theorems*, which enabled one to describe low energy scattering of photons on nucleons even without any underlying theories of strong interactions. These methods relied on exploiting the consequences of gauge-invariance. In another significant work, Low showed how one could calculate Bremsstrahlung of low energy quanta in elementary particle collisions even without the knowledge of any dynamical theory of the structure and interactions of these elementary particles [32]. Again, gauge-invariance played a central role. This work of

low has had a tremendous impact. Nambu and Lurie applied similar ideas to the emission of soft pions [33]. Chirality conservation played the role of gauge-invariance in this work. Precursors to this work were the seminal work by Nambu on Partially conserved axial current [34] and the works by Nambu and Jona-Lasinio on [35] on the dynamical models of elementary particles based on superconductivity.

Some years later Weinberg applied similar techniques to soft-graviton emissions, thereby obtaining the Einstein Principle of Equivalence from first principles. This author in fact applied the same techniques to the problem of gravitational radiation, obtaining thereby an independent derivation of the famous Einstein Quadrupole Formula [36].

Another approach to strong interactions that was highly effective well before QCD was that of *Current Algebras* pioneered by Gell-Mann [37]. The idea behind this is that the non-perturbative S-matrix elements given by the LSZ formalism are often expressed in terms of the matrix elements of *current commutators*. Gell-Mann's essential strategy was to extract these from the Lagrangean of the quark model (he included a neutral vector boson interacting with the quarks), and for the matrix elements use symmetry arguments. An excellent resource for various aspects of Current Algebras that also contains reprints of crucial original papers is the book *Current Algebras* edited by Adler and Dashen [38].

The combination of Current Algebra, PCAC and dispersion theory proved to be very powerful. Among some of its most notable successes are (i) Weinberg's calculation of multiple-pion production [39], (ii) calculation of low energy scattering lengths (scattering amplitudes) [40,41], again by Weinberg, and independently by Y. Tomozawa, and (iii) a calculation of $g_A \equiv \frac{G_A}{G_V}$ by Adler [42] and Weisberger [43], who obtained the values 1.24 and 1.15 respectively. The prevailing experimental value, based on the then prevalent value of 11.7 min for the neutron lifetime, was 1.18 [13]. Over the years, the neutron lifetime has undergone big revisions, the 2019 Particle Data Group (PDG) citing approximately 880s (there is a discrepancy of 8s between free and bottled neutron measurements), translating to a $g_A = 1.2756(11)$ [44]. Based on many of these results, Weinberg and Schwinger pioneered their *Effective Lagrangean* approach, though from very different perspectives. An important basis to both their approaches was the idea of *Chirality* and *Chiral Dynamics* [45,46], and *Partial Symmetry* [47]. By that time Schwinger, the god father of *Operator Field Theory*, had started advocating a renunciation of Operator based techniques, including Current Algebra. Instead, he was promoting his *Source Theory* and its strong phenomenological basis. The reader is referred to his Harvard lectures [48]. He particularly advocated the use of what he called *Numerical Effective Lagrangean Function* [47]. Happily, there was complete convergence between Weinberg's and Schwinger's approaches. We shall return to this towards the end of this section.

Stepping back a little, in 1958 Goldberger and Treiman [49] proved the relation (we have rewritten their Eq. (24) in modern notation)

$$F(0) = \sqrt{2} \frac{g_A M_N}{G_{NN\pi}} \quad (18.22)$$

where $g_A \equiv \frac{G_A}{G_V}$ is as per Eq. (18.8), $G_{NN\pi}$ the pion-nucleon coupling constant, M_N the nucleon mass, and $F(k^2)$ the *Form Factor* for the decay $\pi^- \rightarrow \mu^- + \nu$ defined by the matrix element for this process according to

$$\mathcal{M} = F((p_\mu + p_\nu)^2) \bar{u}(p_\mu) \gamma_5 \gamma_\lambda (1 + \gamma_5) u(p_\nu) (p_\mu + p_\nu)^\lambda \quad (18.23)$$

Equation (18.22), which came to be known as the celebrated *Goldberger–Treiman Relation*, is remarkable in the sense it relates a parameter governing the weak decay of pions to what are clearly parameters of strong interactions. The parameter $F(m_\pi^2)$ can be directly determined in terms of the decay width. In view of the smallness of m_π , $F(0)$ was taken to be close to this. This extrapolation in pion mass was considered to be one of the weaknesses of the Goldberger–Treiman derivation.

Though the momentum $(p_\mu + p_\nu)^\lambda$ can easily be contracted with γ_λ to give $(m_\mu + m_\nu)$ they kept it in this form to reveal some structural aspects of matrix elements of currents. Though only $F(m_\pi^2)$ is relevant to the decay amplitude, they kept this form as their goal was to study $F(k^2)$ through Dispersion Relations. While saturating with intermediate states, they choose to do so with $N\bar{N}$ states instead of the much lighter 3π states, without any justifications. In order to arrive at Eq. (18.22) they had to further demand that their J , which is determined by an integral over some phases, must be larger than about 0.1. For these reasons the derivation of the Goldberger–Treiman relation itself was heavily criticized [50–52] (see also Chap. 12).

Nevertheless, the relation itself was well satisfied by data. We shall first present this relation in its modern form:

$$F_\pi g_{NN\pi} = 2 g_A M_N \quad (18.24)$$

F_π is called the *Pion Decay Constant*, and has the value of 184 MeV. There is much confusion in the literature about its value, with different values of 92, 130 and 184 MeV quoted. All these arise due to different conventions in defining F_π . Of course, with different choices, the Goldberger–Treiman relation also comes in different forms (see [14, 34, 53]).

This led to renewed interest in this relation seeking a better understanding of it. Two very important cornerstones of that development were the proposal of an effective Lagrangean by Gell-Mann and Levy [50] and the proposal of partially conserved axial currents (PCAC) by Nambu [34]. We shall give a simplified and somewhat more direct treatment of both these.

Before doing so, it is worthwhile to take a closer look at the very early calculation of this decay mode by Ruderman and Finkelstein [6], just two years after the discovery of charged pions. The pseudoscalar nature of pions had already been established by then. Because of the presence of neutrinos in the final state, they rightly surmised this to be a weak decay. This was long before the discovery of parity violations and the $V - A$ structure of weak interactions. The generalization of the Fermi theory that was current at that time was to represent the weak Hamiltonian as

$$\mathcal{H}_w = \sum C_i (\bar{\psi}_n \Gamma^i \psi_n) (\bar{\psi}_l \Gamma_i \psi_\nu) \quad H_w = \int \mathcal{H}_w \quad (18.25)$$

where Γ_i are the product of γ -matrices corresponding to scalar (S), pseudoscalar (P), vector (V), axial vector (A) and tensor (T). As commented earlier, it is not necessary to write the hadronic part in terms of nucleon fields, as only matrix elements between nucleon states will appear explicitly, and they are determined by the Lorentz and charge structure of the operators.

The relevant T -matrix element is given by

$$T = \langle e^- \nu | H_w | \pi^- \rangle \quad (18.26)$$

The idea is to introduce completeness relation between the hadronic and leptonic parts, and saturate them accordingly. The vacuum state $|0\rangle$ dominates. The hadronic part then is $\langle 0 | \bar{\psi}_n \Gamma_i \psi_n | \pi^- \rangle$. Though the decay itself is a weak one, this matrix element is to be evaluated by strong dynamics. Parity and Lorentz invariance dictate that only A and P can contribute. To see this, recall that this matrix element in momentum space can only depend on the pion momentum q_μ (say). We refer the reader to [5] for details (Eqs. (29.16)–(29.21)).

For the P and A parts, they use invariance under parity and Lorentz transformations to get

$$\begin{aligned} (2\pi)^{3/2} C_A \langle 0 | \bar{\psi}_p \gamma^\mu \gamma_5 \psi_n | \pi^- \rangle &= i \frac{q^\mu}{m_\pi} f_A(q^2) = i \frac{q^\mu}{m_\pi} f_A(m_\pi^2) \\ (2\pi)^{3/2} C_P \langle 0 | \bar{\psi}_p i \gamma_5 \psi_n | \pi^- \rangle &= i f_P(q^2) = i f_P(m_\pi^2) \end{aligned} \quad (18.27)$$

Little bit of dimensional analysis is helpful here. Both C_A , C_P have mass dimension -2 ; both the axial vector and pseudoscalar operators have mass dim $+3$; the state $|\pi^- \rangle$ has mass dimension -1 (as can be construed from the normalization of one-particle states), and, finally, $|0\rangle$ is dimensionless. Therefore, both matrix elements are dimensionless; factors of m_π have been used to make both f_A and f_P dimensionless.

On using the Dirac equations for the final state particles, it is easy to see that the vacuum dominated T-matrix element takes the form

$$T = -i \frac{\sqrt{m_\nu}}{(2\pi)^{9/2}} \left(\frac{m_l}{m_\pi} f_A + f_P \right) (\bar{u}(p_l) \gamma_5 v(p_\nu)) \quad (18.28)$$

A standard calculation gives the decay rate under the reasonable assumption of very small neutrino masses to be

$$\Gamma(\pi \rightarrow l + \nu) = \frac{m_\pi}{8\pi} \left(1 - \frac{m_l^2}{m_\pi^2} \right)^2 \left(\frac{m_\pi}{m_\pi} f_A + f_P \right)^2 \quad (18.29)$$

This rate is extremely sensitive to f_P . The observed value $(1.25 \pm 0.03) \times 10^{-4}$ of the branching ratio

$$R = \frac{\Gamma(\pi \rightarrow e + \nu)}{\Gamma(\pi \rightarrow \mu + \nu)} = \frac{(\frac{m_e}{m_\pi} f_A + f_P)^2}{(\frac{m_\mu}{m_\pi} f_A + f_P)^2} \frac{(m_\pi^2 - m_e^2)}{(m_\pi^2 - m_\mu^2)} \quad (18.30)$$

determines f_p to be very small. With the advent of parity violation [7,8] and the $V - A$ structure of weak currents [9–11], the issue of f_p was no longer relevant.

Returning to PCAC and the Gell-Mann–Levy effective Lagrangean, we present a simplified, and in our view a more systematic analysis [27]. The Gell-Mann–Levy σ -model started with a non-zero nucleon mass leading to some confusions. We shall instead build this effective Lagrangean step-by-step, turning to the nucleon mass issue at the very end.

Let us start with a discussion of *Vector* (V) and *Axial Vector* (A) currents. The $V - A$ structure of weak interactions is a good motivation. The important questions to understand are their possible conservation and associated invariances, in the spirit of Noether’s theorem. To begin with, we shall look at these currents for *free* nucleons. The electromagnetic current

$$J_{em}^\mu = \bar{\psi}_p(x) \gamma^\mu \psi_p(x) \quad (18.31)$$

is conserved when the free proton obeys the Dirac equation $(i\gamma^\mu \partial_\mu + M_p)\psi_p = 0$. This can be traced to the invariance under global gauge transformations $\delta\psi_p = i\Lambda\psi_p$. In contrast, the charged vector current $\bar{\psi}_p\gamma^\mu\psi_n$ participating in weak interactions is not exactly conserved:

$$\partial_\mu(\bar{\psi}_p\gamma^\mu\psi_n) = i(M_n - M_p)\bar{\psi}_p\psi_n \quad (18.32)$$

But even this current becomes exactly conserved when $M_p = M_n$, which is what happens if isotopic spin is an exact invariance. That motivates the introduction of isovector-vector currents $J_i^\mu = \bar{N}\gamma^\mu \frac{\tau_i}{2} N$ where N denotes the isospinor-spinor nucleon field and $\frac{\tau_i}{2}$ the isospin generators in the doublet representation. The electromagnetic current is obtained when $i = 3$. When isospin invariance is exact, $\partial_\mu J_i^\mu = 0$. This is in accordance with the *Conserved Vector Current* (CVC) hypothesis of Feynman and Gell-Mann [10]. Here we have only shown it for free nucleons. Shortly we shall show it for an interacting theory of pions and nucleons. The power of the CVC hypothesis was that this would hold for arbitrary isospin-invariant theories of strong interactions. As already mentioned, this implied the equality of strengths G_V and G_μ for beta-decay and muon decays.

Now let us turn to the nuclear axial vector current, $\bar{N}\gamma^\mu\gamma_5\frac{\tau_i}{2}N$, and let us assume isospin invariance. It is an elementary exercise to show, for *free* nucleons,

$$\partial_\mu \bar{N}\gamma^\mu\gamma_5\frac{\tau_i}{2}N = 2iM_N\bar{N}\gamma_5\frac{\tau_i}{2}N \quad (18.33)$$

This is far from conserved, even when isospin is an exact symmetry. It can only be conserved for the seemingly unphysical value of $M_N = 0$! However, we shall see that even this seemingly unphysical value has a special significance!

Turning specifically to the nuclear charged axial current $A_\mu^{(+)} \equiv \bar{\psi}_p\gamma_\mu\gamma_5\psi_n$, its matrix elements between nucleon states, which we shall call *direct contribution*, is given by

$$\langle p|A_\mu^{(+)}|n\rangle_{dir} = g_A\bar{u}(k_p)\gamma_\mu\gamma_5u(k_n) \quad (18.34)$$

As already noted, this part of the axial vector is not conserved. Though we explicitly demonstrated it only for free fields, we shall soon see it to be so in an interacting theory too.

This raises the question of whether there is *any* possibility of axial currents being conserved. Considering the perfect symmetry with which axial and vector currents enter weak interactions, and the near equality of G_V , G_A , there is good reason to expect this to extend to their conservations too. A clue to a resolution of this is the simple observation that only the contribution of nucleon fields has been taken into account so far, while in reality the contribution of all strongly interacting (hadronic) fields should be taken into account. Quite naturally, the question of pion's contribution comes to mind. The essence of the Ruderman–Finkelstein calculation [6], and that of Goldberger and Treiman [49] can be captured in the single statement

$$A_\mu^{(+)} = F_\pi \partial_\mu \pi^{(+)} + \dots \quad (18.35)$$

as to the pion's contribution to the axial current. On its own, this would result in a divergence

$$\partial^\mu A_\mu^{(+)} = -F_\pi m_\pi^2 \pi^{(+)} \quad (18.36)$$

upon using the on-shell condition for the pion. In view of the smallness of the pion mass, this lack of conservation of the axial current is negligible in comparison with the non-conservation in the nucleon sector. Nevertheless, this points to a possible exact conservation of the axial current in the limit of vanishing pion mass. Of course, neither the pions nor the nucleons are free. Instead, they interact strongly with each other. So, none of the above may survive a more realistic treatment that includes the pion-nucleon interaction. Here comes the real surprise! Let us consider the pseudoscalar interaction between them

$$\mathcal{L}_{NN\pi} = g_{NN\pi} \bar{\psi}_p \pi^{(+)} \gamma_5 \psi_n \quad (18.37)$$

A generalization of this to the triplet of pions will be given shortly. This interaction leads to what may be called an *indirect* or *pion mediated* contribution to the nuclear matrix elements of the axial current; this can be visualized as the direct coupling of axial current to the pion, followed by a pion propagator, and eventually the interaction of the pion with the nucleons. It is easy to work out that this contribution is

$$\langle p | A_\mu^{(+)} | n \rangle_{med} = i g_{NN\pi} \frac{i F_\pi k_\mu}{k^2 - m_\pi^2} \bar{u}(k_p) \gamma_5 u(k_n) \quad (18.38)$$

with $k_\mu = k_{p\mu} - k_{n\mu}$. The direct and pion-mediated contributions are symbolically displayed in Fig. 18.1, where the dotted lines represent pions:

Thus the total contribution becomes

$$\langle p | A_\mu^{(+)} | n \rangle_{tot} = g_A \bar{u}(k_p) \left\{ \gamma_\mu \gamma_5 - \frac{F_\pi g_{NN\pi} k_\mu}{(k^2 - m_\pi^2)} \gamma_5 \right\} u(k_n) \quad (18.39)$$

Fig. 18.1 Direct and Pion mediated contributions



Consequently,

$$\langle p | \partial^\mu A_\mu^{(+)} | n \rangle = \left\{ 2 g_A M_N - F_\pi g_{NN\pi} \frac{k^2}{(k^2 - m_\pi^2)} \right\} \bar{u}(k_p) \gamma_5 u(k_n) \quad (18.40)$$

In arriving at the above, the full equations of motion for the nucleons interacting with pions have been used. We shall write these down shortly. Here comes the significance of the Goldberger–Treiman relation $2M_N g_A = F_\pi g_{NN\pi}$, which is satisfied by the data $M_N \approx 1000$ Mev, $g_A \approx 1.2$, $F_\pi \approx 184$ Mev, $g_{NN\pi} \approx 13$! On assuming its validity, the expression above reduces to

$$\langle p | \partial^\mu A_\mu^{(+)} | n \rangle = F_\pi m_\pi^2 g_{NN\pi} \frac{1}{k^2 - m_\pi^2} \bar{u}(k_p) \gamma_5 u(k_n) \quad (18.41)$$

Thus we come to the remarkable conclusion that even in the fully interacting theory, the axial current becomes exactly conserved, provided the Goldberger–Treiman relation is satisfied, and the pions are massless! This path-breaking connection was the essential content of Nambu’s PCAC paper [34]! The word *partially conserved* refers to the real-life situation where $m_\pi^2 \neq 0$. To proceed further, we generalize to the iso-triplet of pions:

$$\langle N, p | \partial_\mu \mathbf{A}^\mu | N, p' \rangle = f_\pi g_{NN\pi} m_\pi^2 \frac{1}{k^2 - m_\pi^2} \bar{U}(p) \gamma_5 \boldsymbol{\tau} U(p') \quad (18.42)$$

where $f_\pi \equiv \frac{F_\pi}{2}$ and it is this that takes the value 93 Mev. The change from F_π to f_π is necessitated by the shift to Cartesian components. In so far as single nucleon matrix elements are concerned, this can be expressed equivalently as

$$\partial_\mu \mathbf{A}^\mu = f_\pi m_\pi^2 \boldsymbol{\pi} \quad (18.43)$$

The reader should pay careful attention to the distinction between F_π and f_π . This is the statement of *Partially Conserved Axial Current* (PCAC). We shall now pass on to a description that is not restricted to any particular matrix elements. The full Lagrangean density that has been introduced so far is

$$\mathcal{L}_{\pi N} = \bar{N}(i\gamma^\mu \partial_\mu - M_N)N + \frac{1}{2}[(\partial_\mu \boldsymbol{\pi})^2 + m_\pi^2 \boldsymbol{\pi}^2] - i g_{NN\pi} \bar{N} \gamma_5 \boldsymbol{\tau} \cdot \boldsymbol{\pi} N \quad (18.44)$$

The equations of motion resulting from this Lagrangean are easily written down:

$$\begin{aligned} (i\gamma^\mu \partial_\mu - M_N) N &= i g_{NN\pi} \gamma_5 \boldsymbol{\tau} \cdot \boldsymbol{\pi} N \\ (\partial_\mu \partial^\mu - m_\pi^2) \boldsymbol{\pi} &= -i g_{NN\pi} \bar{N} \gamma_5 \boldsymbol{\tau} N \end{aligned} \quad (18.45)$$

Under isospin transformations (these are the natural generalizations of global electromagnetic gauge transformations)

$$\delta_V N = i \boldsymbol{\Lambda} \cdot \frac{\boldsymbol{\tau}}{2} \quad \delta_V \boldsymbol{\pi} = -\boldsymbol{\Lambda} \times \boldsymbol{\pi} \quad (18.46)$$

with $\boldsymbol{\Lambda}$ a constant. Under these transformations $\mathcal{L}_{\pi N}$ is invariant. Using Noether's first theorem, the corresponding conserved quantities are the vector currents:

$$\mathbf{V}_\mu = \bar{N} \gamma_\mu \frac{\boldsymbol{\tau}}{2} N + \boldsymbol{\pi} \times \partial_\mu \boldsymbol{\pi} \quad (18.47)$$

Using the equations of motion of Eq. (18.45) it is verified that

$$\partial^\mu \mathbf{V}_\mu = 0 \quad (18.48)$$

which is nothing but a statement of Conserved Vector Current (CVC), now for an interacting theory of pions and nucleons.

Turning to axial currents, the corresponding transformations for the nucleon field are $\delta_A N = i \boldsymbol{\omega} \cdot \frac{\boldsymbol{\tau}}{2} \gamma_5 N$, but the axial transformation for pions is not known yet. Instead of guessing this by trial and error, the following easily provable results

$$\delta_A \bar{N} N = \boldsymbol{\omega} \cdot (i \bar{N} \gamma_5 \boldsymbol{\tau} N) \quad \delta_A (i \bar{N} \gamma_5 \boldsymbol{\tau} N) = -\bar{N} N \boldsymbol{\omega} \quad (18.49)$$

will show us a systematic procedure. What these show is that $i \bar{N} \gamma_5 \boldsymbol{\tau} N$ and $\bar{N} N$ transform as a multiplet under axial transformations. To see how this might help decide $\delta_A \boldsymbol{\pi}$, let us consider how the multiplet transforms under the vector transformations:

$$\delta_V (i \bar{N} \gamma_5 \boldsymbol{\tau} N) = -\boldsymbol{\Lambda} \times (i \bar{N} \gamma_5 \boldsymbol{\tau} N) \quad \delta_V (\bar{N} N) = 0 \quad (18.50)$$

In other words, $(i \bar{N} \gamma_5 \boldsymbol{\tau} N)$ transforms exactly like the isovector $\boldsymbol{\pi}$, while $\bar{N} N$ transforms like an isoscalar. Both the axial and vector transformations, taken together, suggest *postulating* a new isoscalar field, say, σ , and the full set of transformation rules to be

$$\begin{aligned} \delta_V N &= i \boldsymbol{\Lambda} \cdot \frac{\boldsymbol{\tau}}{2} N & \delta_A N &= i \boldsymbol{\omega} \cdot \frac{\boldsymbol{\tau}}{2} \gamma_5 N \\ \delta_V \boldsymbol{\pi} &= -\boldsymbol{\Lambda} \times \boldsymbol{\pi} & \delta_A \boldsymbol{\pi} &= -\sigma \boldsymbol{\omega} \\ \delta_V \sigma &= 0 & \delta_A \sigma &= \boldsymbol{\omega} \cdot \boldsymbol{\pi} \end{aligned} \quad (18.51)$$

The reader maybe dissatisfied, and rightly so, that the exact conservation of the axial current is demanding particles not usually considered in strong interactions. Whether a non-observance of such additional particle may make this entire approach worthless. In fact, Gell-Mann and Levy, the creators of this model, had themselves raised this objection; to quote them [50] *The σ -model, although it has some agreeable features, is quite artificial. A new particle is postulated, for which there is no experimental evidence.* They also criticized the model on the grounds of its inability to include strange particles. We seek patience as all such problems (except the ones having to do with strangeness) will be resolved beautifully! Before constructing an exactly chiral and isotopic spin invariant theory, which, by Noether's theorem will guarantee the exact conservation of both vector and axial currents, it is instructive to write down some *invariants* (there are of course, many more):

$$\begin{aligned} I_1 &= \sigma^2 + \pi^2 \\ I_2 &= (\partial_\mu \sigma)^2 + (\partial_\mu \pi)^2 \\ I_3 &= \bar{N} i \gamma^\mu \partial_\mu N \\ I_4 &= \bar{N} N \sigma + i \bar{N} \gamma_5 \tau N \cdot \pi \end{aligned} \quad (18.52)$$

That these are invariants can easily be proved upon using Eq. (18.51).

With the help of these invariants, the following fully invariant Lagrangean density for (N, π, σ) can be constructed:

$$\begin{aligned} \mathcal{L}_{inv} &= \bar{N} i \gamma^\mu \partial_\mu N - g_{NN\pi} \bar{N} (\sigma + i \gamma_5 \tau \cdot \pi) N \\ &+ \frac{1}{2} [(\partial_\mu \sigma)^2 + (\partial_\mu \pi)^2] - \frac{\mu^2}{2} (\sigma^2 + \pi^2) + \frac{\lambda}{4!} (\sigma^2 + \pi^2)^2 \end{aligned} \quad (18.53)$$

The added novelty, apart from invariance under all the transformations of Eq. (18.51), is that this is *renormalizable*. However, the spirit of effective Lagrangeans is not to insist on this but only on their having the same symmetry content of the ultimate theory. An example of such a non-renormalizable, yet fully chiral invariant Lagrangean would be to replace the invariant I_2 by

$$I'_2 = (\square \sigma)^2 + (\square \pi)^2 \quad (18.54)$$

Such Lagrangeans do play a role in the so-called *Chiral Perturbation Theory*.

The equations of motion resulting from $\mathcal{L}_{inv}$ of Eq. (18.53) are

$$\begin{aligned} (\square + \mu^2) \pi + \frac{\lambda}{6} (\sigma^2 + \pi^2) \pi + i g_{NN\pi} \bar{N} \gamma_5 \tau N &= 0 \\ (\square + \mu^2) \sigma + \frac{\lambda}{6} (\sigma^2 + \pi^2) \sigma + i g_{NN\pi} \bar{N} N &= 0 \\ (i \gamma^\mu \partial_\mu - g_{NN\pi} (i \gamma_5 \tau \cdot \pi + \sigma)) N &= 0 \end{aligned} \quad (18.55)$$

Invoking Noether's theorem, the axial and vector currents are found to be

$$\begin{aligned}\mathbf{A}_\mu &= \bar{N} \gamma_\mu \gamma_5 \frac{\boldsymbol{\tau}}{2} N + \sigma \partial_\mu \boldsymbol{\pi} - \boldsymbol{\pi} \partial_\mu \sigma \\ \mathbf{V}_\mu &= \bar{N} \gamma_\mu \frac{\boldsymbol{\tau}}{2} N + \boldsymbol{\pi} \times \partial_\mu \boldsymbol{\pi}\end{aligned}\quad (18.56)$$

It is easy to verify that these currents are exactly conserved upon using the equations of motion of Eq. (18.55).

18.4.1 Group Structure of Chiral Transformations

Using standard methods of Group Theory, it is easy to verify that the vector and axial transformations of Eq. (18.51) form a group, as indeed they should since they leave the Lagrangean invariant. Denoting the *Generators* of isospin and chiral transformations by T_i , X_i respectively, for any field ϕ (this includes N , σ , $\boldsymbol{\pi}$),

$$\delta_V \phi = [\mathbf{A} \cdot \mathbf{T}, \phi] \quad \delta_A \phi = [\boldsymbol{\omega} \cdot \mathbf{X}, \phi] \quad (18.57)$$

the Lie algebra of the group works out to be

$$[T_i, T_j] = i \epsilon_{ijk} T_k; [X_i, X_j] = i \epsilon_{ijk} T_k; [T_i, X_j] = i \epsilon_{ijk} X_k \quad (18.58)$$

It is noteworthy that axial transformations do not form a group on their own. The last of the algebra is a statement that X_i transform like iso-vectors.

The Lie algebra indicates that the group is $O(4)$. But $O(4)$ is isomorphic to the group $SU(2) \times SU(2)$. The latter structure can be made manifest by splitting the nucleon field N into $N_L = \frac{1}{2}(1 - \gamma_5)N$, $N_R = \frac{1}{2}(1 + \gamma_5)N$, and considering *independent* transformations $N_L \rightarrow V N_L$, $N_R \rightarrow U N_R$, $(\sigma + i\gamma_5 \boldsymbol{\tau} \cdot \boldsymbol{\pi}) \rightarrow V(\sigma + i\gamma_5 \boldsymbol{\tau} \cdot \boldsymbol{\pi})U^\dagger$, where U, V are unitary matrices, it can again be seen that $\mathcal{L}_{inv}$ is invariant. This is the $SU(2) \times SU(2)$ structure of the transformations (for details see [53] Sect. 11-4-1). This interpretation sits better with the modern view, basic to both QCD and electro-weak unified theories that the L and R fields are to be thought as independent species. In the early days, it was considered natural to extend this to $SU(3) \times SU(3)$ to include strangeness also, but in the standard model quarks have to be grouped into doublets. Considering that $SU(3)$ and higher extensions are not only not natural but also broken, it makes more sense to stick to the $SU(2) \times SU(2)$ description.

18.4.2 Spontaneous Breaking of Chiral Symmetry

The invariant Lagrangean $\mathcal{L}_{inv}$ is that of the well-known σ -model of Gell-Mann and Levy [50]. They credit the genesis of the model to Schwinger [54]. Gell-Mann and Levy start off with a Lagrangean that we will soon find to be that in the *spontaneously broken phase*, but with suitable change of variables arrive at $\mathcal{L}_{inv}$ except

for a symmetry breaking term of the form $c\sigma$, which we will come to shortly (their σ' -field is exactly the same as our σ -field). Though their manipulations are those of *Spontaneous Symmetry Breaking* they do not explicitly mention this, and to this reason, their model appears contrived at first reading. Their overall normalizations for both the vector and axial currents are a factor 2 different from ours and their axial current has an additional term arising from the explicit symmetry breaking term.

We prefer to start instead from $\mathcal{L}_{inv}$ and proceed systematically first as our primary interest lay in the possibility of the axial vector current being conserved exactly. As it stands, $\mathcal{L}_{inv}$ suffers from the following, rather severe, shortcomings:

- (i) There is no nucleon mass term. Does this corroborate the earlier indication from free nucleons that axial current conservation requires vanishing nucleon mass? However, there is an additional term $g_{NN\pi} \bar{N} N \sigma$ which could play the role of a mass term if σ were a constant, which, unfortunately, it is not!
- (ii) The axial current is conserved even with non-vanishing pion mass! This seems to be against the idea of PCAC according to which the pion mass breaks the conservation of the axial current.
- (iii) The masses of pions and sigma are equal. Phenomenologically no isoscalar with a mass as low as the pion exists. If it did, there would be serious repercussions for nuclear forces.
- (iv) The axial current has no piece that looks like $f_\mu \partial_\mu \pi$ form which had been argued to be the essential content of the works of Ruderman and Finkelstein on the one hand, and, of Goldberger and Treiman on the other. Again, there is a $\sigma \partial_\mu \pi$ piece in the exactly conserved axial current which if $\sigma = f_\pi$ would seem to resolve the difficulty.
- (v) While such a term in the axial current would have required the axial transformation of the pion to have a term like $\delta_A \pi = -f_\pi \omega$, the exact pion transformation of Eq. (18.51) has no such term, and instead, has the full form of $\delta_A \pi = -\sigma \omega$. The two can again be reconciled if $\sigma = f_\pi$, as in the case of the earlier difficulties, but that would violate the σ -transformation law of Eq. (18.51)!

Though $\sigma = f_\pi$ is not an acceptable resolution of the difficulties (i), (iv) and (v) as that would violate the σ -transformation law, something like $\sigma = f_\pi + \tilde{\sigma}$ where $\tilde{\sigma}$ is another *field* would be perfectly acceptable as the σ -transformation law would induce a corresponding transformation law for $\tilde{\sigma}$. In fact, that is precisely what happens when chiral symmetry is spontaneously broken in the same sense in which global gauge-invariance is spontaneously broken in the Ginzburg–Landau description of superconductivity that we have already discussed. We will see that this also resolves the difficulties (ii) and (iii) by making the pion mass vanish by virtue of it becoming the corresponding Goldstone mode. Not surprisingly, the (π, σ) part of $\mathcal{L}_{inv}$ has the same form as the Ginzburg–Landau effective description.

It suffices to look at the spatially homogeneous and temporally constant solutions of the σ, π equations of motion, as those are the ones corresponding to the vacuum solutions. That is (dropping the derivative terms in Eq. (18.55))

$$\begin{aligned} \left\{ \mu^2 + \frac{\lambda}{6} (\sigma^2 + \pi^2) \right\} \pi &= 0 \\ \left\{ \mu^2 + \frac{\lambda}{6} (\sigma^2 + \pi^2) \right\} \sigma &= 0 \end{aligned} \quad (18.59)$$

These are *covariant* under the axial transformations in the sense that the variation of one of them is proportional to the other. Under vector transformations, they are separately *invariant*. Vacuum stability, equivalently a Hamiltonian bounded from below, requires a *positive* λ . Therefore, if μ^2 is also positive, the only solution is

$$\sigma_c = 0 \quad \pi_c = 0 \quad (18.60)$$

This is the chirally symmetric solution in the sense that chiral variations of this solution vanish on this background. With this solution, all the five issues raised before remain unresolved. It is sometimes referred to as the *Wigner realization* of chiral symmetry.

If on the other hand $\mu^2 < 0$, then one can have non-trivial solutions

$$\sigma_c^2 + \pi_c^2 = \frac{6|\mu^2|}{\lambda} \quad (18.61)$$

Even though $\sigma_c = \pi_c = 0$ are still solutions, they represent *unstable* solutions as the vacuum energy is higher for them. This is completely analogous to the situation in the Ginzburg–Landau theory. Unlike the symmetric solution, which is unique, now there are a continuum of solutions, each related to another through a chiral transformation. But the ground state itself has to be unique and only one of the infinitely degenerate solutions has to be selected, just as in the case of superconductivity. This represents a spontaneous breakdown of chiral symmetry. In the case of superconductivity, it was a global U(1) that was spontaneously broken, but now it is the global $SU(2) \times SU(2)$ that is spontaneously broken. Since isospin should remain an invariance, the $SU(2) \times SU(2)$ is spontaneously broken to the *diagonal* $SU(2)$. This is the subgroup of chiral transformations with $\omega = \Lambda$.

Chiral transformations can be used to bring any solution to the form $\sigma_c = \sqrt{\frac{6|\mu^2|}{\lambda}} \equiv f_\pi$, $\pi_c = 0$. The *potential* $V(\sigma, \pi)$ can then be arranged as

$$V(\sigma, \pi) = \frac{\lambda}{4!} (\sigma^2 + \pi^2 - f_\pi^2)^2 - \frac{\lambda f_\pi^4}{4!} \quad (18.62)$$

Thus the symmetric solution corresponds to $V_{sym} = 0$ while the spontaneously broken, sometimes called the asymmetric solution, corresponds to $V_{ssb} = -\frac{\lambda f_\pi^4}{4!}$. Therefore, the asymmetric solution, lower in energy, is the stable solution. Now to understand all the excitations, one introduces $\sigma = f_\pi + \tilde{\sigma}$, $\pi = \tilde{\pi}$. The physical fields whose vacuum values are zero are $\tilde{\sigma}$, $\tilde{\pi}$.

Remarkably, in the spontaneously broken phase, all the five difficulties get resolved at once. All the kinetic terms maintain their form in terms of the new fields. We take up the rest one by one:

- (i) **Nucleon Mass** What was earlier $g_{NN\pi} \bar{N} \sigma N$ now becomes $g_{NN\pi} f_\pi \bar{N} N + g_{NN\pi} \bar{N} \tilde{\sigma} N$. This implies a nucleon mass $M_N = g_{NN\pi} f_\pi$. Recalling that $f_\pi = \frac{F_\pi}{2}$ this looks like the Goldberger–Treiman relation, but the factor g_A is missing. In other words, to this order $g_A = 1$. We shall discuss the implications of this shortly.
- (iv) The $\sigma \partial_\mu \pi$ in the axial current now becomes $f_\pi \partial_\mu \pi + \tilde{\sigma} \pi$; the first of these resolves this difficulty.
- (v) The pion transformation law now becomes $\delta_A = -f_\pi \omega - \tilde{\sigma} \omega$ so this difficulty also gets resolved.
- (ii) and (iii): upon substituting for σ , it is easy to see that the pion and sigma equations of motion become

$$\begin{aligned} \square \pi + \frac{\lambda f_\pi}{3} \tilde{\sigma} \pi + \frac{\lambda}{6} (\tilde{\sigma}^2 + \pi^2) \pi &= 0 \\ \square \tilde{\sigma} + \frac{\lambda f_\pi^2}{3} \tilde{\sigma} + \dots &= 0 \end{aligned} \quad (18.63)$$

The first of these says that the pion mass in this phase is exactly zero. This is just a consequence of the Goldstone Theorem, exactly as in the case of superconductivity. However, there only one Goldstone boson was produced whereas now we have three of them, together forming the iso-triplet. In the former, it was U(1) symmetry with one generator that was spontaneously broken. Now $SU(2) \times SU(2)$ with six generators spontaneously breaks to the diagonal $SU(2)$ (isospin invariance) with 3 unbroken generators. Therefore the number of broken generators is three which equals the number of Goldstone bosons. This restores the earlier view that the axial current can be exactly conserved only for zero pion mass.

- **Interaction Terms:** The $\tilde{\sigma}$ squared mass, which was tachyonic to begin with $-|\mu^2|$ has now become physical with squared mass taking the positive value of $\frac{\lambda f_\pi^2}{3}$ and the degeneracy in masses has also disappeared, removing the phenomenological objections.

In conclusion, pions are Goldstone bosons of spontaneous breaking of Chiral Symmetry. Curiously, Heisenberg was the first to think of pions as the massless particles associated with the spontaneous breakdown of isospin invariance; however, that would have yielded scalar pions in contradiction to their observed pseudoscalar nature.

Having satisfactorily clarified the issue of masses, we now write down the interaction terms in the SSB phase:

$$\mathcal{L}_{int}^{ssb} = -g_{NN\pi} \bar{N} (\tilde{\sigma} + i\gamma_5 \boldsymbol{\tau} \cdot \boldsymbol{\pi}) N + \frac{\lambda f_\pi}{6} \tilde{\sigma} (\tilde{\sigma}^2 + \pi^2) + \frac{\lambda}{4!} (\tilde{\sigma}^2 + \pi^2)^2 \quad (18.64)$$

A consequence of this is that the physical sigma particle is unstable against decay into two pions. Depending on λ , the width could be very large, making it difficult to detect such a particle. The other important consequences of this interaction Lagrangean are the so-called *Chiral Cancellations*, which we shall address shortly.

18.4.3 Non-linear Realization of Chiral Symmetry

So far we discussed a *linear* realization of chiral symmetry which necessitated the chiral multiplet (σ, π) in the sense that the chiral transformations giving rise to the chiral invariant theory were linear transformations among these. The price one had to pay for this was the introduction of the σ . This, as mentioned before, could be a source for some phenomenological discomfort. Rather remarkably, there is a *non-linear* realization of chiral symmetry wherein the σ can be eliminated resulting in a theory with only nucleons and pions, and yet be chiral invariant. In a more precise sense, a theory of pions and nucleons invariant under transformations whose generators X_i would satisfy $[X_i, X_j] = i \epsilon_{ijk} T_k$. Even if an isoscalar particle with broad width is found, the non-linear realization would still make sense for phenomena with all momenta very small in comparison with the sigma mass.

The basis for obtaining the non-linear realization is the invariant character of $\sigma^2 + \pi^2$ under the linear transformations. Therefore, the sigma can be eliminated on imposing the invariant restriction $\sigma^2 + \pi^2 = f_\pi^2$. That is, by $\sigma = \sqrt{f_\pi^2 - \pi^2}$. Apart from the low momentum restriction mentioned before, this reduction makes sense only as long the pion field does not exceed f_π in magnitude.

This elimination, not surprisingly, is self-consistent in the sense that it preserves the $\delta_A \sigma = \omega \cdot \pi$ when $\delta_A \pi = -\sigma \pi$ is used. Upon elimination of σ , the pion chiral transformation law becomes (the isospin transformation remains unchanged as it does not involve the σ)

$$\delta_A \pi = -\sqrt{f_\pi^2 - \pi^2} \omega \quad (18.65)$$

which is clearly a non-linear transformation law. Expressed in terms of the generators X_i :

$$[X_i, \pi_j] = i \sqrt{f_\pi^2 - \pi^2} \delta_{ij} \quad (18.66)$$

In an elegant and powerful paper Weinberg laid out the theory of the most general non-linear realization of chiral symmetry [55]. He showed they are of the type

$$[X_i, \pi_j] = i \delta_{ij} f(\pi^2) + \pi_i \pi_j g(\pi^2) \quad (18.67)$$

The Jacobi Identity

$$[X_k, [X_i, \pi_j]] + [\pi_j, [X_k, X_i]] + [X_i, [\pi_j, X_k]] = 0 \quad (18.68)$$

along with $[X_i, X_j] = i \epsilon_{ijk} T_k$; $[T_i, \pi_j] = i \epsilon_{ijk} \pi_k$ imposes the restriction

$$f g - 2 f' (f + \pi^2 g) = 1 \quad (18.69)$$

The transformation we obtained upon eliminating the σ corresponds to the choice $g = 0$, $f(\pi^2) = -\sqrt{f_\pi^2 - \pi^2}$. Different choices of (f, g) correspond to different

ways of redefining the pion field. These redefinitions do not affect the observable S-matrix elements.

It is clear that upon eliminating the σ -field as above, the only relevant pieces of the pion Lagrangean left are (the rest are all constants)

$$\mathcal{L}_\pi = \frac{1}{2} \left\{ (\partial\pi)^2 + \frac{(\pi \cdot \partial_\mu \pi)^2}{f_\pi^2 - \pi^2} \right\} \quad (18.70)$$

It is to be noted that this is no longer *renormalizable*, but it has all the symmetries of the renormalizable σ -model. That indeed is in the spirit of effective descriptions.

Weinberg [55] suggested to write this in the form

$$\mathcal{L}_\pi = \frac{1}{2} (\mathcal{D}_\mu \pi)^2 \quad (18.71)$$

where

$$\mathcal{D}_\mu \pi = \partial_\mu \pi + \frac{\partial_\mu \pi^2}{4(f_\pi^2 - \pi^2 + \sqrt{(f_\pi^2 - \pi^2)})} \pi \quad (18.72)$$

was called the *Chiral Pion Covariant Derivative* by Weinberg. It is a covariant derivative in the same sense as the covariant derivatives of General Theory of Relativity or Gauge Theories, i.e. they transform the same way as the fields they are covariant derivatives of. In this particular instance they transform under the non-linear chiral transformations exactly as the pions. This is more than just an analogy as the different choices of (f, g) can be thought of as being related by general coordinate transformations where the pion fields act as the coordinates.

We conclude this subsection by writing down the pion-nucleon chiral invariant, isospin-invariant Lagrangean density (for brevity only the first two terms in the expansion of σ are shown):

$$\mathcal{L}_{\pi N} = \bar{N}(i\gamma^\mu \partial_\mu - M_N) + \frac{1}{2} (\mathcal{D}_\mu \pi)^2 - g_{NN\pi} \bar{N} \left(-\frac{\pi^2}{2f_\pi} + i\gamma_5 \boldsymbol{\tau} \cdot \boldsymbol{\pi} \right) N \quad (18.73)$$

18.4.4 Chiral Cancellations

In the spontaneously broken phase there are important cancellations in the amplitudes for low momentum scattering phenomena. These are very general and can in fact be understood purely as consequences of the pions being Goldstone bosons. They happen for all systems in which symmetries are spontaneously broken.

Let us first consider $\pi\pi$ -scattering, and to bring out the nuances clearly, let us first consider the symmetric phase where the pions are not Goldstone bosons. To leading order, only the $\lambda \frac{\pi^4}{4!}$ term contributes and the scattering amplitude is $A(\pi\pi \rightarrow \pi\pi) \approx i\lambda$.

But in the spontaneously broken phase, there are additional interactions of the type $\frac{\lambda f_\pi}{6} \tilde{\sigma} (\tilde{\sigma}^2 + \pi^2)$ (see Eq. (18.64)). On using that $m_\sigma^2 = \frac{\lambda f_\pi^2}{3}$, it is seen that for momenta much smaller than this mass, the first contribution gets almost fully cancelled by the $\tilde{\sigma}$ -exchange contributions, leaving a residual contribution vanishing as $\simeq \frac{k^2}{f_\pi^2}$. In the non-linear realization this is even more direct to see as the term responsible for this scattering is $\frac{1}{8 f_\pi^2} (\partial_\mu \pi^2)^2$. It is noteworthy that the residual amplitude is completely *independent* of λ !

Next, we consider pion-nucleon scattering. In the symmetric phase, this scattering in leading order arises at the second order perturbation in the interaction $-i g_{NN\pi} \bar{N} \gamma_5 \boldsymbol{\tau} \cdot \boldsymbol{\pi} N$ and at very low momentum takes the constant value of $\propto \frac{g_{NN\pi}^2}{M_N}$. But in the spontaneously broken phase, there is a $\tilde{\sigma}$ exchange contribution which cancels most of it, leaving a small residual amplitude also vanishing as $\frac{k}{f_\pi}$. This cancellation depends both on the $\tilde{\sigma}$ -mass as well as the analog of the Goldberger–Treiman relation $g_{NN\pi} f_\pi = M_N$ (note the absence of g_A). In the non-linearly realized theory, there is an additional interaction $g_{NN\pi} \bar{N} \frac{\pi^2}{2 f_\pi} N$ leading to the cancellation. Here too, the residual amplitude is independent of λ . The reader is urged to go through the worked Example 5.4 in [56] to see the details. Thus at very low momenta, the pion-nucleon interaction, instead of being very strong, is actually very weak. Chiral cancellations formed the basis for model-independent perspectives on the scalar-isoscalar channel and multinucleon forces in Nuclear Physics, proposed by the author in collaboration with Anishetty and Sharatchandra [57]. They were studied by us in the broader context of Goldstone Bosons in [58].

18.4.5 Finite Pion Mass

In actuality, pions do have mass, and therefore cannot be Goldstone bosons strictly. However, their mass being so low, they are *almost* Goldstone bosons, and for this reason are called *Pseudo-Goldstone Bosons*. So how should one incorporate the finite pion mass into the scheme of spontaneous breaking of chiral symmetry? One could think of adding an explicit symmetry breaking term like $-\frac{m_\pi^2}{2} \pi^2$. But the pion field redefinition freedom would imply that such a procedure is not unique. In other words, one could equally well have added terms like $\frac{m_\pi^2}{f_\pi^2} \pi^4$, etc. (see the remarks by Weinberg [55]). A way of introducing symmetry breaking with some degree of control would be to add a term that has definite transformation properties. That suggests adding a term like $f_\pi m_\pi^2 \sigma$ instead. That would also maintain the renormalizability of the σ -model, even though, as we shall see shortly, one should not be insisting on renormalizability for effective descriptions. The choice of symmetry breaking has a profound effect on such observables as scattering lengths, which we shall discuss in some depth later.

Adding such a symmetry breaking term would lead again to PCAC as formulated earlier:

$$\partial_\mu \mathbf{A}^\mu = f_\pi m_\pi^2 \pi \quad (18.74)$$

In QCD, such a symmetry breaking has to be formulated in terms of quarks and gluons. It is taken to be $m_q \bar{\psi}_q \psi_q$. Consistency of the two descriptions requires

$$\langle 0 | f_\pi m_\pi^2 \sigma | 0 \rangle = f_\pi^2 m_\pi^2 = m_q \langle \bar{\psi}_q \psi_q \rangle \quad (18.75)$$

which is nothing but the famous Gell-Mann–Oakes–Renner relation [59] (see [14], Eq. (19.7.16) for a SU(3) extension).

18.4.6 Phenomenological Lagrangeans

Is the Gell-Mann–Levy σ -model an effective description of strong interactions, even at low momenta? The answer is both “yes” and “no”! The answer is yes because it is formulated in terms of observable strongly interacting fields like pions and nucleons whereas QCD, the currently accepted microscopic theory of strong interactions is formulated in terms of the unobservable quark and gluon fields (more on this in the next chapter). Yet, it cannot be called a phenomenological effective description because something as basic as the Goldberger–Treiman relation is obtained only with $g_A = 1$ while experimentally it is more like 1.2.

One of the suggestions is that the σ -model is still to be treated as an *Operator Field Theory* and that $g_A = 1$ is only the *bare* value, and the renormalized value may be closer to the observed value. The Adler–Weisberger derivation lends some credence to this point of view. In fact Gell-Mann and Levy themselves took this stand and discussed some broad aspects of renormalization in [50] itself. A fairly detailed account of the renormalization of the σ -model can be found in Sect. 11-4-2 of [53].

Both Julian Schwinger and Steven Weinberg advocated a next level of effective description which the former called *Numerical Lagrange Function* [46,47], and the latter *Phenomenological Lagrangean* [45,60] where a non-operator treatment of the fields, i.e. the tree-level amplitudes would adequately describe the experimentally observed aspects. Weinberg was motivated by his own scattering length calculations [40] for both low energy pion-pion scattering as well as pion-nucleon scattering. Schwinger also gave an analysis of these scattering lengths. Subsequent to the pioneering works by Schwinger and Weinberg, many important developments took place through the works of Sidney Coleman, Julius Wess, Bruno Zumino and others. The reader is referred to the review of early developments by Gasiorowicz and Geffen [61].

Even to this day effective field theories, equivalently, phenomenological Lagrangeans, are at best a poor man’s theory to many. It is interesting that the two pioneers, Weinberg and Schwinger, themselves had strong differences of opinion (as mentioned in Ref. 7 of [55])! To Schwinger, they provided a suitable area for the study of symmetries, as long as the origins of symmetries remained obscure [62],

while to Weinberg, it was a matter of uneasiness to use symmetries at a phenomenological level when it was not clear how *any* fundamental Lagrangean could give rise to the supposed symmetry! To this author, phenomenological Lagrangeans are like the light at the end of the tunnel, which, while not as dazzlingly bright as the fundamental Lagrangeans would be, nevertheless offer hopes for systematically getting there. In fact the role played by chiral symmetry in the eventual formulation of QCD is a case in point.

Let us see how these phenomenological effective Lagrangeans work by first considering the pion scattering lengths. As clearly stated at length by Weinberg in [55], chiral symmetry alone is not sufficient to fix the real-life scattering lengths; the pattern of symmetry breaking by the finite pion masses is also crucial. In the same paper, Weinberg has a table showing how the pion-pion scattering length depends on various choices of symmetry breaking. So the spirit of the effective Lagrangeans for $\pi - \pi$ scattering is to take them to be the terms quartic in pion fields arising from Eq. (18.70), and the symmetric breaking term $f_\pi m_\pi^2 \sqrt{f_\pi^2 - \pi^2}$:

$$\mathcal{L}_{\pi\pi}^{eff} = \frac{1}{2 f_\pi^2} (\pi \cdot \partial_\mu \pi)^2 - \frac{m_\pi^2}{8 f_\pi^2} (\pi^2)^2 \quad (18.76)$$

Weinberg's description (see his Chap. 19 of [14]) is somewhat different, being based on the pseudo-vector approach (this is also the approach favoured by Schwinger [48]). Taking his symmetry breaking (Eq. (19.5.34) of [14] along with his symmetric effective Lagrangean of his Eq. (19.5.23)), what one has (after due allowance made for the metric conventions) is

$$\mathcal{L}_{\pi\pi}^{wein} = -\frac{1}{F_\pi^2} \partial_\mu \pi \cdot \partial^\mu \pi \pi^2 + \frac{m_\pi^2}{2 F_\pi^2} (\pi^2)^2 \quad (18.77)$$

It should be recalled that $F_\pi = 2 f_\pi$. The difference in form of these two effective Lagrangeans is due to the difference in choice of the chiral covariant derivatives, which can also be understood as due to pion field redefinitions. It is straightforward to verify that the two differ by an irrelevant four-divergence.

It is easy to appreciate Weinberg's statement about the importance of the symmetric breaking term. In fact, an accurate determination of the scattering lengths would be a good way to fix the symmetry breaking term (see also Table-I of [55]). Weinberg deduces the contribution to the $\pi_a(p_A) + \pi_b(p_B) \rightarrow \pi_c(p_C) + \pi_d(p_D)$ scattering arising from the symmetric term to be (in his metric convention)

$$\begin{aligned} \mathcal{M}_{abcd}^{sym,w} &= 4 F_\pi^{-2} \{ \delta_{ab} \delta_{cd} (-p_A \cdot p_B - p_C \cdot p_D) + \delta_{ac} \delta_{bd} (p_A \cdot p_C + p_B \cdot p_D) \\ &\quad + \delta_{ad} \delta_{bc} (p_A \cdot p_D + p_B \cdot p_C) \} \end{aligned} \quad (18.78)$$

This is Eq. (19.5.25) of [14]. However, he incorrectly concludes that, in terms of the Mandelstam invariants $s = -(p_A + p_B)^2$, $t = -(p_A + p_C)^2$, $u = -(p_A + p_D)^2$ (his defining equations after his Eq. (19.5.26)), this equals

$$\mathcal{M}_{abcd} = 4 F_\pi^{-2} \{ \delta_{ab} \delta_{cd} (s - m_\pi^2) + \delta_{ac} \delta_{bd} (t - m_\pi^2) + \delta_{ad} \delta_{bc} (u - m_\pi^2) \} \quad (18.79)$$

whereas it should be

$$\mathcal{M}_{abcd}^{sym,w} = 4 F_\pi^{-2} \{ \delta_{ab} \delta_{cd} (s - 2 m_\pi^2) + \delta_{ac} \delta_{bd} (t - 2 m_\pi^2) + \delta_{ad} \delta_{bc} (u - 2 m_\pi^2) \} \quad (18.80)$$

This would of course give the wrong amplitude as it does not satisfy the *Adler Consistency Condition* which states that the amplitude should vanish if any three of the pions are on mass shell while the fourth has zero four-momentum (unphysical) [63]. In other words, the amplitude should vanish at $s = t = u = m_\pi^2$. The resolution of course comes from taking into account the contributions from the symmetry breaking (this must clearly be an oversight in his book, as he was the one to stress the importance of symmetry breaking for scattering lengths):

$$\mathcal{M}_{abcd}^{breaking,w} = 4 F_\pi^{-2} m_\pi^2 \{ \delta_{ab} \delta_{cd} + \delta_{ac} \delta_{bd} + \delta_{ad} \delta_{bc} \} \quad (18.81)$$

which restores the total contribution to Eq. (18.79), which indeed satisfies the Adler consistency condition. It may seem curious that the consistency condition only obtains for a particular choice of symmetry breaking, but it is not surprising if one recalls that the chief ingredient in the derivation of these consistency conditions was the form of PCAC that is tied to the pattern of symmetry breaking!

The same total expression obtains from Eq. (18.76) also, but the symmetric and breaking terms are organized differently. In fact,

$$\mathcal{M}_{abcd}^{sym} = f_\pi^{-2} \{ \delta_{ab} \delta_{cd} s + \delta_{ac} \delta_{bd} t + \delta_{ad} \delta_{bc} u \} \quad (18.82)$$

while the breaking term contribution is

$$\mathcal{M}_{abcd}^{breaking} = -f_\pi^{-2} m_\pi^2 \{ \delta_{ab} \delta_{cd} + \delta_{ac} \delta_{bd} + \delta_{ad} \delta_{bc} \} \quad (18.83)$$

The scattering lengths are calculated by dividing $\mathcal{M}$ at threshold, i.e. $s = 4 m_\pi^2, t = u = 0$ by a normalizing factor of $32 \pi m_\pi$. The threshold value of $\mathcal{M}_{abcd}$ is $\mathcal{M}_{abcd}^{thresh} = 4 m_\pi^2 F_\pi^{-2} [3 \delta_{ab} \delta_{cd} - \delta_{ac} \delta_{bd} - \delta_{ad} \delta_{bc}]$. On using $\mathcal{M}_{abcd}^{(0)} = \frac{1}{3} \delta_{ab} \delta_{cd}$, and, $\mathcal{M}_{abcd}^{(2)} = \frac{1}{2} (\delta_{ac} \delta_{bd} + \delta_{ad} \delta_{bc} - \frac{2}{3} \delta_{ab} \delta_{cd})$ for the isospin channels $T = 0$ and $T = 2$, the corresponding scattering lengths a_0, a_2 can be worked out to be

$$a_0 = \frac{7 m_\pi}{8 \pi F_\pi^2} \approx 0.16 m_\pi^{-1} \quad a_2 = -\frac{m_\pi}{4 \pi F_\pi^2} \approx -0.046 m_\pi^{-1} \quad (18.84)$$

The experimental values quoted in [14] (Ref. [28], circa 1992) are $m_\pi a_0 = 0.26 \pm 0.05$, $m_\pi a_2 = -0.028 \pm 0.012$. The agreement is not spectacular. Weinberg claims higher order corrections in $\frac{m_\pi}{2\pi F_\pi}$ seem to improve agreement.

The role of the symmetry breaking term can be further elucidated by introducing a free parameter β to characterize it (such an exercise was also carried out by Schwinger and is explained in [48]) in Eq. (18.83):

$$\mathcal{M}_{abcd}^{breaking} = 4 \beta F_\pi^{-2} m_\pi^2 \{ \delta_{ab} \delta_{cd} + \delta_{ac} \delta_{bd} + \delta_{ad} \delta_{bc} \} \quad (18.85)$$

It is easy to see that this modifies the scattering lengths to

$$a_0 = \frac{m_\pi}{4\pi F_\pi^2} \left(1 + \frac{5\beta}{2}\right) \quad a_2 = \frac{m_\pi}{4\pi F_\pi^2} (-2 + \beta) \quad (18.86)$$

Quite clearly the scattering lengths a_0, a_2 can be changed at will by changing the symmetry breaking parameter β . However, the combination $(2a_0 - 5a_2)$ does not depend on β at all and is given entirely by the chiral symmetric parts. This was the conclusion of both Schwinger and Weinberg. However, only the choice $\beta = 1$ satisfies the Adler consistency condition.

Now we turn to a discussion of the phenomenological Lagrangeans for the pion-nucleon system. Once again, one possibility is to interpret the σ -model itself as one such, with only its tree diagrams taken into account. But as already mentioned earlier, this runs into difficulties with $g_A = \frac{G_A}{G_V}$ which comes out to be exactly 1 instead of the phenomenological value close to 1.2. Even on theoretical grounds, the Adler–Weisberger derivation of this is closer to this observed value.

This can be resolved on noting that the phenomenological Lagrangeans need not necessarily be *renormalizable* unlike the σ -model. Weinberg advocates the addition of non-renormalizable interactions which are nevertheless chiral invariant, such as his Eq. (19.5.50) [14]:

$$\mathcal{L}'_{\pi N} = i g' \bar{N} \left\{ \gamma^\mu \gamma_5 \left(\frac{\boldsymbol{\tau}}{2} \cdot \boldsymbol{\pi} \partial_\mu \sigma - \sigma \partial_\mu \left(\frac{\boldsymbol{\tau}}{2} \cdot \boldsymbol{\pi} \right) \right) + \frac{\boldsymbol{\tau}}{2} \cdot (\boldsymbol{\pi} \times \gamma^\mu \partial_\mu \boldsymbol{\pi}) \right\} N \quad (18.87)$$

This adds to $g_A = 1$ a term linear in g' . Thus the value of g_A can be made to be anything in a chirally symmetric way. In the spirit of phenomenological Lagrangeans, g' is adjusted to agree with the experimental values.

We refer the reader to [14], Sect. 19.5, for the details of the pion-nucleon S-wave scattering amplitudes. Ignoring the g' terms, and hence corrections due to deviation of g_A from 1, Weinberg finds the threshold amplitude to be

$$\mathcal{M}_{ba} = -4i \frac{m_\pi}{F_\pi^2}, \mathbf{t} \cdot [\mathbf{t}^{(\pi)}]_{ba} \quad (18.88)$$

where $\mathbf{t} = \frac{\boldsymbol{\tau}}{2}$ and $\mathbf{t}^{(\pi)}$ is the pion isovector matrix. The S-wave scattering lengths for the total isospin $T = \frac{1}{2}, \frac{3}{2}$ turn out to be

$$a_{1/2} = \frac{m_\pi}{\pi F_\pi^2 \left(1 + \frac{m_\pi}{M_N}\right)} \quad a_{3/2} = -\frac{m_\pi}{2\pi F_\pi^2 \left(1 + \frac{m_\pi}{M_N}\right)} \quad (18.89)$$

Numerically, $a_{1/2} \approx 0.15 m_\pi^{-1}$, and, $a_{3/2} \approx -0.075 m_\pi^{-1}$. The experimental values quoted by Weinberg are, respectively, $m_\pi a_{1/2} = 0.173 \pm 0.003$, $m_\pi a_{3/2} = -0.101 \pm 0.004$. The agreement is fairly good. Weinberg's values obey $a_{1/2} + 2a_{3/2} = 0$. This is due to his neglect of the g' terms. Both the issues of corrections

due to the g' terms and the $g_A \neq 1$ can be seen in a much simpler, but equivalent approach presented by Schwinger in [48]. We shall present a broad overview of the same.

The starting point of Schwinger is his Eq. (3.7) for the interaction terms for low energy $\pi - N$ system

$$\mathcal{L}_{\pi N}^{sch} = \frac{f}{m_\pi} \bar{N} i \gamma^\mu \gamma_5 \boldsymbol{\tau} \cdot \partial_\mu \boldsymbol{\pi} N + \left(\frac{f_0}{m_\pi} \right)^2 \bar{N} \gamma^\mu \boldsymbol{\tau} \cdot (\partial_\mu \boldsymbol{\pi} \times \boldsymbol{\pi}) N \quad (18.90)$$

Like Weinberg's modified interactions, Schwinger also introduces two distinct couplings and non-renormalizable terms. Working out the exact correspondence between the approaches is tedious, involving partial integrations, pion field redefinitions, etc., which we shall not attempt. Schwinger fixes his f -coupling by data on low energy P-wave scattering to be $f = 1.01 \pm 0.01$.

Schwinger then goes on to calculate the low energy S-wave pion-nucleon scattering lengths. His results are

$$\begin{aligned} m_\pi (2a_{3/2} + a_{1/2}) &= -3 \frac{f^2}{4\pi} \frac{m_\pi}{M_N} \approx -0.03 \\ m_\pi (a_{3/2} - a_{1/2}) &= -3 \frac{f_0^2}{2\pi} \end{aligned} \quad (18.91)$$

He too finds that only f_0 terms would yield $2a_{3/2} + a_{1/2} = 0$. But having determined f from P-wave data, he considers both the couplings. Using the observed values he fixes $f_0 = 0.84 \pm 0.03$.

He then goes on to consider the vector and axial transformations and the associated currents in Sect. 3.2 of [48], based on his 1967 work [47]. He arrives at the expression

$$\mathbf{j}_\mu = \partial_\mu \boldsymbol{\pi} \times \boldsymbol{\pi} + \frac{m_\pi}{2f_0} \partial_\mu \boldsymbol{\pi} + \bar{N} \frac{\boldsymbol{\tau}}{2} \left(1 - i\gamma_5 \frac{f}{f_0} \right) N + \dots \quad (18.92)$$

In his own words, this puts together in one package the CVC, the Goldberger–Treiman relation, and the value $g_A = \frac{f}{f_0} = 1.2$. Yet, only what he called *Numerical Lagrange Functions* or what Weinberg called *Phenomenological Lagrangeans* are employed.

18.4.7 Chiral Perturbation Theory

So the spirit of such effective Lagrangeans is to allow anything that conforms to what is thought to be the underlying symmetry, even a partial realization of it such as PCAC, without imposing such restrictions as renormalizability, etc. At one point of time, even as recently as the formulation of electro-weak unification and QCD, renormalizability was considered to be an essential and desirable feature of fundamental theories. But with the advent of Wilson's Renormalization Group-based approaches,

this is no longer the thinking. In their own ways, Schwinger and Feynman had also come to similar views. In that sense, at every level of knowledge, our theories are always effective theories only.

A point of caution is in order at this point. What one thinks of as non-renormalizable is an *ultraviolet* point of view, and typically, higher derivative interactions grow faster with momentum scale and cause more severe divergence difficulties at higher energies. But the same terms can also be viewed from an *infrared* perspective too. From that perspective, higher the number of derivatives in an interaction, less important will it be at *low* energies. Therefore, in the context of low energy effective descriptions, the most dominant would be those that are quadratic in momenta for Bosonic d.o.f or linear for Fermionic d.o.f, and as we go to higher momentum scales, interactions with increasing number of derivatives start becoming more important, and it would appear as if there is a need to include non-renormalizable interactions. In reality, it is just a *Taylor Expansion* in low momenta.

Let us start with pions first. Then the non-linear model with $\mathcal{L} = \frac{1}{2} \mathcal{D}\pi^2$ would be the dominant description, where $\mathcal{D}$ is the chiral covariant derivative (that this is already a non-renormalizable interaction may be confusing at first sight!) That can be avoided by working in the linearly realized model, where this would be $\frac{1}{2}[(\partial_\mu\sigma)^2 + (\partial_\mu\pi)^2]$. As we move to higher momenta, this will have to be augmented by terms like $(\mathcal{D}_\mu\mathcal{D}_\mu)^2$ and $(\mathcal{D}_\mu\mathcal{D}_\nu)^2$, whose coefficients are now dimensionful full (in the linearly realized version these would be, for example, $(\square\sigma)^2 + (\square\pi)^2$). As just explained, these higher derivative terms can be viewed as both increasingly non-renormalizable terms or as higher order terms in the Taylor expansion for low momenta.

Then it would appear that effective descriptions to arbitrarily desired accuracy would be obtained by writing down terms up to a certain number of derivatives and treat their coefficients as parameters to be phenomenologically determined. Even though the number of free parameters has increased, these will still have a lot of predictive power, even if used only for their tree diagrams. That turns out to be too restricted a point of view and hides one of the essential aspects of effective Lagrangeans. That is, effective descriptions come with them the range of momenta they are valid for. Therefore, fluctuations over that range are always present and must be accounted for.

The programme of *Chiral Perturbation Theory* aims to do this. But most practitioners of this write down Feynman diagrams and carry out integrations over loop momenta to arbitrarily high values. They then carry out a formal expansion in external momenta. Some of the divergences (as well as convergent terms) so encountered are absorbed into counter terms which are of the same form as the naive infrared expansion already alluded to. But one also obtains *universal* features that have non-polynomial dependence (typically logarithmic) on momenta (pion mass included), which the naive infrared expansions would have missed. This area has been very active for several decades and it is hard to do full justice to it with a few references, but sources with good technical and conceptual coverages are the Schladming lectures by Gasser [64], Les Houches lectures by Golterman [65], by Manohar [66]. Additionally, the reader is referred to [67–69].

But such integrations over arbitrarily large loop momenta miss the essential spirit of the infrared expansions underlying the effective descriptions. In fact, the leading order description is that of an *infrared Gaussian fixed point* in the language of the renormalization group (RG), and extending to marginally larger values of momenta should be achieved through small perturbations around the fixed point. Such a procedure should also uncover the non-polynomial corrections. This important aspect was forcefully brought out by Weinberg in [60]. The infrared fixed point structure was also highlighted by us in [58].

The situation for pion-nucleon interactions seems rather different. As we already saw, even gross features like $g_A \neq 1$ require going to higher orders. This is sometimes ascribed to the fact that even in the low momentum region, there is a large mass scale like that of the nucleon mass. Weinberg has made several perceptive remarks on this thorny issue in [60].

18.4.8 Anomalous Sector

In retrospect, one of the most embarrassing failures of early current algebra was its inability to account for the $\pi^0 \rightarrow \gamma\gamma$ -decay! The early treatments showed the decay amplitude to be highly suppressed in comparison to the observed rate. This was pointed out independently by Veltman [70] as well as by Sutherland [71]. Rather remarkably, Schwinger had given a successful account of this decay as early as 1951 in his classic paper on Gauge Invariance [72]! This conflict is what is referred to as *Anomalies*. This forms a very important topic in QFT, with wide ramifications. The reader is referred to the extensive treatment of Anomalies in Weinberg's [14] (the entire Chap. 22), as well as in [53]. Both these explain the Veltman–Sutherland puzzle and how anomalies resolve it. In the specific context of this book, the central extension of the Virasoro algebras of the dual model as well as the dual strings is also an anomaly.

The traditional explanation for anomalies has been in terms of the impossibility of regularizing QFT's in such a way that all the symmetries present at the classical level are maintained at the quantum level. In the context of the π^0 -decay, the conflict is between the *global* chiral symmetry and the *local* gauge-invariance. The 1951 work of Schwinger had chosen to regularize maintaining gauge-invariance. The theme was resurrected in 1969 by the works of Adler [73], and, Bell and Jackiw [74]. A highly non-trivial result about Anomalies was the work of Adler and Bardeen, also in 1969 [75], which showed that Anomalies are unaffected by higher order radiative corrections. In the path-integral quantization of QFT's, this deep result is a consequence of the fact that only the path-integral measure accounts for the entire anomaly. This was shown by Fujikawa [76]. This is explained clearly in [14].

Returning to the neutral pion decay, the naive form of PCAC gets modified to

$$\partial^\mu A_\mu^{(3)} = f_\pi m_\pi^2 \pi^0 - \frac{\alpha}{8\pi} \epsilon_{\mu\nu\rho\sigma} F^{\mu\nu} F^{\rho\sigma} \quad (18.93)$$

where α is the fine-structure constant. This is called the *Abelian Anomaly* as the only gauge fields involved are the Abelian electromagnetic fields. The concept of anomalies gets generalized when there are non-Abelian gauge fields. That is a very rich and fascinating topic. Weinberg gives a clear and detailed exposition in [14]. We now briefly discuss how anomalies affect the effective Lagrangeans. The startling implication of Eq. (18.93) is that even when chiral symmetry violations are absent, for example, when $m_\pi = 0$, the axial current is still not conserved. Noether's theorem can be used to cast this as

$$\frac{\delta_{A_3} \mathcal{L}_{eff}^{anom}}{\delta \omega_3} = -\frac{\alpha}{8\pi} \epsilon_{\mu\nu\rho\sigma} F^{\mu\nu} F^{\rho\sigma} \quad (18.94)$$

The *Anomalous Effective Lagrangean* is obtained by so to say “integrating” this equation. Modulo the anomalous effective Lagrangean, the effective Lagrangeans are chiral symmetric or partially chiral symmetric, and their constructions are along the lines already discussed. The first such anomalous effective action was derived by Aviv, Hari Dass and Sawyer [77] and predicted a whole new class of reactions with photons and *odd* number of pions. While their results for the sector involving only neutral pions was correct, the rest had issues with gauge-invariance. Those defects were resolved soon in [78–80]. Of these [80] only addressed the reactions $\gamma \rightarrow 3\pi$, $\gamma\gamma \rightarrow 3\pi$. Wess and Zumino gave very powerful conceptual and technical tools for the construction of anomalous effective Lagrangeans. But their methods only work for groups like $SU(3) \times SU(3)$, but not for $SU(2) \times SU(2)$ as that is a so-called *Anomaly-Free Group*. More technically, the *Wess–Zumino Term* is absent for them. However this author showed how anomalous effective Lagrangeans for an arbitrary number of pions can be found in closed form for abelian anomalies in $SU(2) \times SU(2)$ with electromagnetism [79], using the powerful methods of Weinberg's non-linear realizations of chiral symmetry [55]. These cannot be obtained by the methods of [78], nor of a more recent one by Witten [81]. All these nuances are discussed in the work by Golterman and Hari Dass [82]. In that work QCD-inspired effective actions were proposed for pions and vector mesons, including the weak vector bosons, as well as extensions into the anomalous sector.

References

1. L.D. Landau, E.M. Lifshitz, *Theory of Elasticity* (Butterworth-Heinemann, 2005)
2. H. Jeffreys, B.S. Jeffreys, *Methods of Mathematical Physics* (Chap. 3). Cambridge University Press
3. J.D. Jackson, *Classical Electrodynamics*, 2nd edn. Wiley Eastern Ltd
4. E. Fermi, *Z. f. Phy.* **88**, 161 (1934)
5. S. Gasiorowicz, *Elementary Particle Physics*. John Wiley and Sons
6. M. Ruderman, R. Finkelstein, *Phys. Rev.* **76**, 1458 (1949)
7. C.N. Yang, T.D. Lee, *Phys. Rev.* **104**, 254 (1956)
8. C.S. Wu et al., *Phys. Rev.* **105**, 1413 (1957); *Phys. Rev.* **106**, 1361 (1957)
9. E.C.G. Sudarshan, R.E. Marshak, *Phys. Rev.* **109**, 1860 (1958)
10. R.P. Feynman, M. Gell-Mann, *Phys. Rev.* **109**, 193 (1958)

11. J.J. Sakurai, *Nuovo Cimento* **7**, 649 (1958)
12. N. Cabibbo, *Phys. Rev. Lett.* **10**, 531 (1963)
13. A.N. Sosenovskiy et al., *Nuc. Phys.* **10**, 395 (1959); C.P. Bhalla, *Phys. Lett.* **19**, 691 (1966)
14. S. Weinberg, *Quantum Theory of Fields-II*. Cambridge University Press
15. F.J. Hasert et al., *Phys. Lett.* **46**, 121 (1973)
16. W. Meissner, R. Ochsenfeld, *Naturwissenschaften* **21**(44), 787 (1933)
17. J. Bardeen, L.N. Cooper, J.R. Schrieffer, *Phys. Rev.* **108**, 1175 (1957)
18. V.L. Ginzburg, L.D. Landau, *JETP (USSR)* **20**, 1064 (1950)
19. M.R. Beasley, *Notes on the Ginzburg-Landau Theory*. ICMR School on Novel Superconductors, UCSB, 2–15 August 2009
20. S. Weinberg, *Prog. Theor. Phys. Suppl.* No **86**, 43 (1986)
21. Y. Nambu, *Phys. Rev.* **117**, 648 (1960)
22. P.W. Anderson, *Phys. Rev.* **130**, 439 (1963)
23. J. Goldstone, *Nuovo Cimento* **9**, 154 (1961)
24. J. Goldstone, A. Salam, S. Weinberg, *Phys. Rev.* **127**, 965 (1962)
25. L.N. Cooper, *Phys. Rev.* **104**, 1189 (1956)
26. F. London, H. London, *Proc. Roy. Soc. A* **149**(866), 71 (1935)
27. N.D. Hari Dass, *Three Lectures on Chiral Symmetry*. [arXiv:hep-ph/0506169v1](https://arxiv.org/abs/hep-ph/0506169v1) (hep-ph)
28. Y. Nambu, *Symmetry Breakdown and Small Mass Bosons*, *Fields and Quanta* 1 (1970), p. 33; Reprinted in *Broken Symmetry*, Selected Papers of Y. Nambu, ed. by T. Eguchi, K. Nishijima. World Scientific
29. *Broken Symmetry*, Selected Papers of Y. Nambu, ed. by T. Eguchi, K. Nishijima. World Scientific
30. F.E. Low, *Phys. Rev.* **96**, 1428 (1954)
31. M. Gell-Mann, M.L. Goldberger, *Phys. Rev.* **96**, 1433 (1954)
32. F.E. Low, *Phys. Rev.* **110**, 974 (1958)
33. Y. Nambu, D. Lurie, *Phys. Rev.* **125**, 1429 (1962)
34. Y. Nambu, *Phys. Rev. Lett.* **4**, 380 (1960)
35. Y. Nambu, G. Jona-Lasinio, *Phys. Rev.* **122**, 345 (1961); *Phys. Rev.* **124**, 246 (1961)
36. N.D. Hari Dass, *Low Energy Theorems for Gravitational Radiation*. Niels Bohr Institute Preprint, NBI-HE-81-45
37. M. Gell-Mann, *Physics* **1**, 63 (1964)
38. *Current Algebras*, ed. by S. Adler, R.F. Dashen. Benjamin (1968)
39. S. Weinberg, *Phys. Rev. Lett.* **16**, 879 (1968)
40. S. Weinberg, *Phys. Rev. Lett.* **17**, 616 (1966)
41. Y. Tomozawa, *Nuovo Cimento* **46A**, 707 (1966)
42. S.L. Adler, *Phys. Rev.* **140**, B736 (1965)
43. W.I. Weisberger, *Phys. Rev.* **143**, 1302 (1965)
44. W.J. Marciano, A. Sirlin (2018). [arXiv:1802.01804](https://arxiv.org/abs/1802.01804) [hep-th]
45. S. Weinberg, *Phys. Rev. Lett.* **18**, 188 (1967)
46. J. Schwinger, *Phys. Lett. B* **24**, 473 (1967); *Phys. Rev.* **167**, 1432 (1968)
47. J. Schwinger, *Phys. Rev. Lett.* **18**, 923 (1967)
48. J. Schwinger, *Particles and Sources* (Gordon and Breach, 1969)
49. M.L. Goldberger, S. Treiman, *Phys. Rev.* **110**, 1178 (1958)
50. M. Gell-Mann, M. Levy, *Nuovo Cimento* **16**, 705 (1960)
51. M. Gell-Mann, *Phys. Rev.* **125**, 1067 (1962)
52. R.F. Sawyer, *Phys. Rev.* **116**, 231 (1959)
53. C. Itzykson, J.-B. Zuber, *Quantum Field Theory*. McGraw Hill Publishers
54. J. Schwinger, *Ann. Phys.* **2**, 407 (1957)
55. S. Weinberg, *Phys. Rev.* **166**, 1568 (1968)
56. B. de Wit, J. Smith, *Field Theory in Particle Physics*, vol. I. North Holland Personal Library
57. R. Anishetty, N.D. Hari Dass, H.S. Sharatchandra, *Implications of Chiral Symmetry for Scalar-Isoscalar Channel and Multinucleon Forces* (1996). [arXiv:9612306](https://arxiv.org/abs/9612306) [hep-ph]
58. R. Anishetty, R. Basu, N.D. Hari Dass, H.S. Sharatchandra, *Int. J. Mod. Phys. A* **14**(22), 3467 (1999)

59. M. Gell-Mann, R.J. Oakes, B. Renner, *Phys. Rev.* **175**, 2195 (1968)
60. S. Weinberg, *Physica A* **96**(1–2), 327 (1979)
61. S. Gasiorowicz, D.A. Geffen, *Rev. Mod. Phys.* **41**, 531 (1969)
62. J. Schwinger, *Phys. Rev.* **152**, 1219 (1966)
63. S.L. Adler, *Phys. Rev.* **139**, B1638 (1965)
64. J. Gasser, *Schladming Lectures* (2003). [arXiv:0312367](https://arxiv.org/abs/0312367) [hep-ph]
65. M.F.L. Golterman, *Les Houches Lectures* (2009). [arXiv:0912.4042](https://arxiv.org/abs/0912.4042)
66. A. Manohar (2018). [arXiv:1804.05863](https://arxiv.org/abs/1804.05863)
67. J. Gasser, H. Leutwyler, *Phys. Rep.* **87**, 77 (1982)
68. B. Ananthanarayan, *Eur. Phys. J. ST* **231**, 291 (2022)
69. B. Ananthanarayan, *Pramana* **61**, 911 (2003)
70. M. Veltman, *Proc. Roy. Soc. A* **301**, 107 (1967)
71. D.G. Sutherland, *Nuc. Phys. B* **2**, 433 (1967)
72. J. Schwinger, *Phys. Rev.* **82**, 604 (1951)
73. S.L. Adler, *Phys. Rev.* **177**, 2426 (1969)
74. J.S. Bell, R. Jackiw, *Nuovo Cimento* **60A**, 47 (1969)
75. S.L. Adler, W.A. Bardeen, *Phys. Rev.* **182**, 1517 (1969)
76. K. Fujikawa, *Phys. Rev. Lett.* **42**, 1195 (1979)
77. R. Aviv, N.D. Hari Dass, R.F. Sawyer, *Phys. Rev. Lett.* **26**, 591 (1971)
78. J. Wess, B. Zumino, *Phys. Lett. B* **37**, 95 (1971)
79. N.D. Hari Dass, *Phys. Rev. D* **5**, 1542 (1972)
80. S.L. Adler, B.W. Lee, S.B. Treiman, A. Zee, *Phys. Rev. D* **4**, 3497 (1971); R. Aviv, A. Zee, *Phys. Rev. D* **5**, 2372 (1972); M. Terentev, *JETP Lett.* **14**, 140 (1971)
81. E. Witten, *Nuc. Phys. B* **223**, 422 (1983)
82. M.F.L. Golterman, N.D. Hari Dass, *Nuc. Phys. B* **277**, 739 (1986)



Quantum Chromodynamics (QCD)—A RQFT for Strong Interactions

19

19.1 Introduction

In Chap. 3, we discussed the failures in arriving at a Quantum Field Theory of nuclear forces to describe the pion-nucleon interactions. Among various sources for these failures, apart from the large couplings rendering the perturbative approaches useless, was the proliferation of new particles over and above the pions and nucleons. This led to a trend away from quantum field theories, and towards the S-matrix approaches, eventually leading to a String Theory of hadrons. We have elaborated this path in the earlier chapters. That the string theory required what appeared to be unphysical space-time dimensionalities, tachyonic states, and excitations that obviously were not part of the hadronic world like massless excitations with spin-1 and spin-2, also undermined confidence in them as the correct descriptions of strong interactions. But even through all this, the embers of QFT were not wholly dead and Murray Gell-Mann, who contributed so strongly to the analytic S-matrix program, was one who was a strong believer in the usefulness of QFT. In this chapter, we describe the remarkable resurgence of RQFT as a candidate for describing strong interactions. Rather remarkably, the growth of these powerful ideas happened around the same time as the final phases of the dual resonance models and string theory.

We give a short, but self-contained introduction to the development of Quantum Chromodynamics, or, QCD in short, as the RQFT of the theory of strong interactions. In doing so, I have drawn heavily from the excellent reviews by Harald Fritzsch, one of the creators of QCD along with Murray Gell-Mann [1, 2]. Fritzsch passed away last month (16 August 2022) and we dedicate this chapter to his memory. There are also excellent interviews with Gell-Mann on QCD (see, for example, the interview by Norman Dombey [3]).

QCD is a vast area of current research and the aim of this chapter is not to describe all its wonderful details. Instead, we shall highlight those aspects that are directly relevant for the rest of the book. To that extent, only the pure-gauge aspects will be dealt with.

19.2 Historical Backgrounds to QCD

Let us start with the proliferation of particles that more or less ended the quests for meson field theories. This proliferation consisted of the so-called *resonances* in pion-nucleon scattering, as well as some totally new kinds of particles. Both these classes will play big roles in this narrative. Of the second category are particles that came to be called *Strange Particles*. The Λ^0 hyperon was among the first such to be discovered. It was neutral, spin-1/2, with a mass $M_{\Lambda^0} = 1115.44 \text{ MeV}$, considerably heavier than the nucleons (protons and neutrons) with average masses around 939 MeV. It was an isospin-singlet, i.e. $I = 0$. Also discovered were the strange mesons K^+ , K^0 , $\bar{K}^0$, K^- with masses around 495 MeV. The reader is referred to the book by Gasiorowicz for a wealth of information about the new particles and their theoretical relevance [4].

19.2.1 Strangeness

A puzzling feature of particles like K-mesons and Λ^0 -hyperons was that together they could be produced very fast, but individually their decays took considerably longer. It was as if different interactions governed their productions and decays, at total variance with one's theoretical understandings of production and decays. It was Abraham Pais who made a breakthrough with his proposal of *Associated Production* [5]. According to this, certain combinations of these particles (for example, $\Lambda^0 + K^+$ or $K^+ + K^-$) could be produced very fast but individually they would take very long to decay. From Pais's proposal, Gell-Mann came up with the concept of *Strangeness* quantum number S , with the stipulation that it is conserved by strong interactions but not the weak interactions. He [6], and Nakano and Nishijima [7] generalized an earlier formula relating charge and isospin to

$$Q = I_3 + \frac{B + S}{2} \quad (19.1)$$

This quantified the concept of strangeness and suddenly there was a great transparency about the new particles. Some of the strangeness quantum number assignments were: $\Lambda^0 (S = -1)$, $K^+ (S = 1)$, $K^0 (S = 1)$, $\bar{K}^0 (S = -1)$, etc. A particularly exciting and elegant outcome was the work of Pais and Gell-Mann on the quantum mechanics of the K^0 , $\bar{K}^0$ -system [8].

19.2.2 Sakata Model

Attempts continued to bring some order into the rapid proliferation of particles: one direction was to see if all the particles could be envisaged as being “made out” of a small set of particles, the old idea of *compositeness*. Sakata [9] proposed the *Sakata Model* which initially appeared to bring some order. His idea was that all particles

could be made of a basic triplet $\tilde{p}, \tilde{n}, \tilde{\Lambda}^0$ which had the same quantum numbers as p, n, Λ^0 in so far as their charge, isotopic spin and strangeness were concerned. For example, the physical π^+ can be thought of as being composed of $(\tilde{p}\tilde{\bar{n}})$, the K^+ as $(\tilde{p}\tilde{\Lambda}^0)$, etc. Likewise, the baryons can be envisaged as $p \simeq (\tilde{p}\tilde{p}\tilde{\bar{p}})$, etc. But attempts at viewing the basic triplet as **3** representation of a SU(3) soon ran into many difficulties, the main source of the difficulties being that anti-particles had to be used to build both mesons and baryons. This arose because Sakata had assigned baryon number $B = 1$ to the triplet. Another consequence of this was that minimum strangeness for Baryons would be -2 , and maximum $+1$, soon to be contradicted.

19.2.3 The Eightfold Way

At this stage Gell-Mann [10] and Neeman [11] proposed another way out, called the *Eightfold Way*! By this time the known baryons formed an octet, i.e. (p, n) with $I = 1/2, S = 0$, $(\Sigma^+, \Sigma^0, \Sigma^-)$ with $I = 1, S = -1$, Λ^0 with $I = 0, S = -1$, and (Ξ^-, Ξ^0) with $I = 1/2, S = -2$. The known mesons also formed an octet, i.e. (π^+, π^0, π^-) with $I = 1, S = 0$, (K^+, K^0) with $I = 1/2, S = 1$, $(\bar{K}^0, K^-)$ with $I = 1/2, S = -1$, and η^0 with $I = 0, S = 0$. Gell-Mann and, independently Neeman conjectured that these octets belong to the **8** representation of SU(3). This single hypothesis paved the way for much progress!

This immediately led to a consideration of the resonances in πN -scattering. A prominent resonance, already established in early 1950s (see [4] for a detailed discussion of this resonance) was the so-called Δ -resonance. At a mass of 1238 MeV, it was established to be a $J = 3/2$ with isospin $I = 3/2$. The isospin assignment came from the fact that this very prominent resonance was also seen in $\pi^+ p$ -scattering. Now, if eightfold way was right, it had to belong to one of the SU(3) representations in the group decomposition:

$$\mathbf{8} \otimes \mathbf{8} = \mathbf{1} \oplus \mathbf{8} \oplus \mathbf{8} \oplus \mathbf{10} \oplus \mathbf{10}^* \oplus \mathbf{27}$$

Because of its isospin $I = 3/2$, the only possibilities were **10** or **27**. Of these **10** was the strongly favoured assignment. That would imply that the other partners of this resonance, all with $J = 3/2$, must consist of a $I = 1, S = -1$, $I = 1/2, S = -2$, and $I = 0, S = -3$. The last would have $Q = -1$ from the Gell-Mann-Nishijima formula. This was called Ω^- and was a prediction of the eightfold way. Its mass had been predicted to be 1679 MeV based on broken SU(3) considerations. The big triumph came when it was experimentally discovered in 1964, with a mass of 1675 MeV! [12]. Sometimes in literature this prediction is credited to the *Quark Model*, which we shall discuss next. But it is really a prediction of the Eightfold Way (I am myself guilty of this in [13]. There, my recounting the history of QCD was also incomplete, which I am correcting here!) It is worth pointing out that such a state would have been impossible in the Sakata model.

19.3 The Quark Models

19.3.1 Gell-Mann-Zweig Quark Models

After the formulation of the Eightfold way and the underlying $SU(3)$ structure, a question naturally occurred to many, including Gell-Mann himself. That was the puzzle as to why the fundamental representation $\mathbf{3}$ of $SU(3)$ played no role at all! Of course, in the Sakata model this representation did play a role, but owing to the quantum number assignments that model ran into a number of difficulties. The answer to this puzzle was provided independently by Murray Gell-Mann and George Zweig, in the form of their Quark and Ace models, respectively, [14–16]. It is amusing that George Zweig, at that time a graduate student, never published his results in a journal!

The dramatic departure of the Quark Model from the Sakata model was in the quantum number assignments. The baryon numbers of the quarks were chosen to be $B = 1/3$, and the Gell-Mann-Nishijima charge formula was assumed to hold good nevertheless. This led to the electric charge assignments $(2/3, -1/3, -1/3)$ for the so-called u , d , s quarks (modern notation) when their strangeness was taken to be $(0, 0, -1)$. Now all the baryons were taken to be made of three quarks. Unlike the Sakata model, no antiquarks went into the making of baryons. This is summarized as the *Charge-matrix* in Fig. 19.1.

The vertical columns of this matrix are the values of (I_3, Y, Q) and the horizontal rows are the values for the (u, d, s) quarks; the hypercharge Y is given by $Y = B + S$. Fractional values of electric charge and baryon number did raise eyebrows, there was no real inconsistency if these bizarre values were only assigned to hitherto unobserved (as it turned out, to remain unobserved!) building blocks.

The $SU(3)$ group multiplications

$$\mathbf{3} \otimes \bar{\mathbf{3}} = \mathbf{1} \oplus \mathbf{8} \quad \mathbf{3} \otimes \mathbf{3} \otimes \mathbf{3} = \mathbf{1} \oplus \mathbf{8} \oplus \mathbf{8} \oplus \mathbf{10} \quad (19.2)$$

naturally gave rise to the meson and baryon octets, as well as the crucial decuplet of baryons which contained the Δ resonances and the Ω^- !

In what follows we shall use *Hypercharge* $Y = B + S$ in place of strangeness, without any loss of generality. Whether the quarks were really the constituents of all hadrons was far from clear. The model was seen by some as merely a book-keeping device for the classification and properties of all hadrons. The (u, d) quarks were taken to be $I = 1/2$ while s quarks were $I = 0$, $S = -1$. The $SU(3)$ was clearly a *broken* symmetry as showed by the considerable differences between strange and non-strange baryons and mesons.

Fig. 19.1 The charge-matrix for Gell-Mann-Zweig quark model

$$\begin{array}{ccc} \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \end{array}$$

19.3.2 The Statistics Difficulties

The decuplet which played such an important role in supporting the Eightfold way posed serious problems for the Gell-Mann-Zweig quark model. This had to do with the fact that Δ^{++} had to be composed of three identical quarks, i.e. (*uuu*). Likewise, the Ω^- , the star of the Eightfold way, had to be composed of three identical quarks too, i.e. (*sss*). The spins of all the decuplet states being $J = 3/2$, were totally symmetric under spin-exchange. Quarks being Fermions, one had to conclude that the spatial wavefunctions had to be *antisymmetric*, in both cases. In other words, the orbital angular momenta had to be non-zero. The general expectations were that such non-zero orbital angular momentum states, because of the angular momentum barriers, had to be of higher energy.

A word of caution is necessary here. It is not totally inconceivable that a complicated dynamics would result in states of higher angular momentum with lower energies. For example, it could be caused by very strong spin-orbit interactions, etc. But in the present context, there appear to be other reasons for favouring angular momenta to be zero such as the magnetic moment calculations (see [17] p.30).

Greenberg [18] attempted to solve the problem on the basis of his *Parastatistics* which is neither Fermi-Dirac statistics nor Bose-Einstein. We shall not go into further details of this.

19.3.3 Han-Nambu (HN) Quark Model

Soon after, in 1965, Han and Nambu came up with a remarkable proposal to solve the spin-statistics problem. The concept of *Colour* was introduced in this paper [19] though the authors did not use that phraseology. The essential idea was to increase the number of quarks from three to nine, more precisely from a single triplet of the Gell-Mann-Zweig model to *three* triplets. But there were many other radical departures. The group structure was now expanded to $SU(3) \otimes SU(3)$ with one of them, as in the Gell-Mann-Zweig model, identified with *flavour* and the other, with new quantum numbers that could be called *colour*. In essence, each flavour of quark came in three types. By requiring states to be singlets (totally antisymmetric) under the new $SU(3)$, the statistics problem was immediately averted.

Han and Nambu wanted their quarks to have *integral* electric charges and hypercharges. Nevertheless, they wanted to maintain the Gell-Mann-Nishijima relation between Q , I_3 , Y . Though this is reminiscent of the Sakata model, they still wanted the baryons to be made up of three quarks, without any antiquarks. Their paper presents a systematic search for triplets based on $U(3)$. Their requirements of integral charges and hypercharges inevitably led them to a mixing of flavour and colour. To see this more clearly, let us take a look at their “charge-matrices”; for each of them, the vertical columns are (I_3, Y, Q) , while the horizontal rows denote the flavour types, i.e. (*u*, *d*, *s*) (Fig. 19.2): We have called the three colours, Red, Green and Blue; that’s not the terminology Han and Nambu used. More precisely, colour is the **3** representation of the colour $SU(3)$. It is at once obvious that there is a *colour-flavour*

Fig. 19.2 Charge matrices
for Red, Green, Blue colours
of Han-Nambu model

$$\begin{array}{ccc|ccc|ccc} \frac{1}{2} & -\frac{1}{2} & 0 & 0 & -1 & -\frac{1}{2} & 1 & 0 & \frac{1}{2} \\ 1 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & -1 & -1 & 1 & 0 & 0 \end{array}$$

mixing. In particular, the electric charge Q depends both on flavour and colour. To see this, consider, for example, the u-quark; when its colour is red, its charge is $Q = 1$, while it is $Q = 0$ for green, and $Q = 1$ for blue. This colour-flavour mixing will turn out to pose many difficulties for the Han-Nambu model. One may wonder how the electric charge could depend on colour and at the same time the Gell-Mann-Nishijima formula be valid! This happens because I_3 and Y are also colour dependent in this scheme.

19.3.4 π^0 -Decay as a Test for Quark Models

With two competing quark models that were also radically different, interest naturally arose as to find empirical tests that would favour one or the other. The earliest of such tests was the neutral-pion decay into two photons. A legitimate doubt that would arise is as to how two different models can be tested with a single test as both models have various parameters! The key to this is to express the π^0 -decay in terms of as many observables as possible. It turns out, due to the approximate chiral symmetry of low-energy pion phenomena, that the pion mass m_π and the *pion decay constant* f_π are two such. In terms of them, the decay rate can be expressed as

$$\Gamma(\pi^0 \rightarrow \gamma\gamma) = \frac{m_\pi^3}{32\pi f_\pi^2} \left(\frac{\alpha}{\pi}\right)^2 (\sum Q_+^2 - \sum Q_-^2)^2 \quad (19.3)$$

where α is the fine-structure constant and $Q_\pm$ are the charges of the $I_3 = \pm \frac{1}{2}$ components that couple to the pion. An exhaustive account of the neutral-pion decay can be found in the review [20]. The reader is also referred to [21] for a clear discussion.

The early calculations of this decay rate had used the nucleon doublet for which $Q_P = 1$, $Q_n = 0$ yielding $Q_p^2 - Q_n^2 = 1$ resulting in a value 7.87 eV for Γ as against the exptal value 7.95. The reader is referred to the paper of Julian Schwinger [22] for a discussion of some of the initial theoretical difficulties and their resolution. Adler compared the predictions of the fractional and integral quark models [23], and found that the Han-Nambu model in fact agrees with the exptal value much better. What is worse, the Gell-Mann-Zweig model predicts a rate that is $\frac{1}{9}$ the correct value!

In the Han-Nambu model,

$$(1_u^2 - 0_d^2) + (1_u^2 - 0_d^2) + (0_u^2 - 1_d^2) = 1 \quad (19.4)$$

while in the Gell-Mann-Zweig model

$$(Q_u^2 - Q_d^2) = \frac{4}{9} - \frac{1}{9} = \frac{1}{3} \quad (19.5)$$

we will soon see a striking resolution of this!

19.3.5 Colour in Gell-Mann-Zweig Model

In 1972 Bardeen, Fritzsch and Gell-Mann came up with a radically different solution to the spin-statistics problem [24]. The novelty of their solution is best appreciated by looking at their “charge-matrices” shown in Fig. 19.3 and contrast them with Fig. 19.2:

There is no longer any flavour-colour mixing! The electric charge, I_3 and Y are all independent of colour. This will be seen to have profound consequences for quark dynamics. We will discuss this shortly. This colour scheme solves the statistics problem by requiring all baryons to be singlets under a colour SU(3), which is now considered to be an *exact* symmetry. Of course, there is no dynamical explanation as yet as to why only colour-singlet states are observable. But the Han-Nambu scheme also had this issue. Eventually this will be seen to be the *Quark Confinement Problem*.

Apart from solving the statistics problem, the new colour scheme automatically solves the π^0 -decay discrepancy also as the amplitude for the decay is now tripled. Another test of the quark charges is the so-called R-ratio:

$$R = \frac{\sigma(e^+e^- \rightarrow \text{hadrons})}{\sigma(e^+e^- \rightarrow \mu^+\mu^-)} = \frac{\sum_i Q_i^2}{Q_\mu^2} \quad (19.6)$$

at high energies. Including s , d , u quarks with three colours, one gets 2 for this. Experimental results in the summer of 1972 indeed showed this ratio to be above 2. With further increase of $\sqrt{s}$ for the e^+e^- -system, this should keep going up. With (u, d, s, c, b) quarks this should reach 3.66. Such a behaviour has indeed been seen and forms an important test for colour.

This naturally raises the possibility of using the same R -ratio to test the Han-Nambu model. An elementary calculation yields $R = 4$ when only the u , d , s quarks are included. This contradicts the experimental values by a huge margin! Does this immediately rule out the Han-Nambu integrally charged quarks? Surprisingly, the answer turns out to be a NO! This subtlety will be clarified shortly.

Fig. 19.3 The charge matrices for fractionally charged quark model with colour

$\frac{1}{2}$	$-\frac{1}{2}$	0	$\frac{1}{2}$	$-\frac{1}{2}$	0	$\frac{1}{2}$	$-\frac{1}{2}$	0
$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{2}{3}$
$\frac{2}{3}$	$\frac{2}{3}$	$-\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$-\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$-\frac{1}{3}$
$\frac{2}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$\frac{2}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$\frac{2}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$

19.4 Towards Theories of Quark Dynamics

While the quark models provided a picture that proved to be very good as a “book-keeping” device for the new proliferation of particles, it was by no means clear whether one should take the quarks more seriously, and, in particular, what would be the nature of the dynamics governing their behaviour. For our purposes, we shall only concentrate on the dynamical models related to the Han-Nambu quarks on the one side, and the Gell-Mann-Fritzsch quarks on the other. What is amazing is that to this day there are no “acid tests” to distinguish these two approaches that seem so different theoretically. While Han and Nambu addressed the question of dynamics in their very first paper introducing integrally charged quarks (with colour) [19], Fritzsch and Gell-Mann took nearly 8 years (in 1972 with W.A. Bardeen) to address the same for fractionally charged quarks (also with colour).

What needs to be emphasized is that the path to this dynamics as indicated in most textbook treatments of QCD is to render *local* the *global* invariances, introducing the *gauge fields* for the purpose. Even the creators of QCD, i.e. Fritzsch and Gell-Mann, as we shall see, did not take that path to start with, though Han and Nambu did suggest that. That recipe has not always worked for all global invariances either. Take isotopic spin for example; the local gauge-invariance approach for this as suggested by Schwinger [25] and Sakurai [26] was not successful.

19.4.1 Han-Nambu Approach

Rather remarkably, Han and Nambu remarked on the dynamical aspects of their quark model in their first paper itself. But these remarks do not reveal the details of the dynamics they had in mind. However, they explicitly mention an *octet* of coloured gluons as mediators of the quark dynamics. They also consider these gluons to be *flavour singlets*. The remarks are mostly verbal in nature, and are to be found in the fourth para after their Eq. (27). They do talk about the gauge bosons being massive pointing to a breakdown of the colour SU(3). Details of how the masses are generated are only implicit. Around the time of this paper, the so-called Higgs effect to generate masses for gauge vector bosons was already known and Nambu himself had played a major role in those important theoretical developments. So it is entirely plausible that he had such mechanisms in mind, but nothing is really explicitly stated.

In his 1966 paper on *A systematics of Hadrons in Subnuclear Physics* [27], Nambu reiterates his dynamical scheme of an octet of coloured gluons, transforming as the 8-dimensional adjoint representation of colour SU(3) (described by Nambu as “coupled to infinitesimal generators of the group”) coupled to the colour SU(3) *triplet* of quarks with universal coupling g . This is explicitly stated after Eq. (19) of the paper. He goes on to mention various reasons for breaking this symmetry, but again the dynamical details are very sketchy. Fritzsch, in [1], states further that in the 1966 paper Nambu *gauged* both the colour and flavour SU(3)’s, and that there was a mixing between the gauge bosons of the two sectors. Even this is not clearly spelt out in the 1966 paper, though these statements are true as we shall see soon.

Despite the fact that in gauging the Han-Nambu quark theory, one had to necessarily break both the flavour and colour gauge invariances, and despite the fact that the two sets of gauge bosons had to necessarily mix because of the colour-flavour mixing demanded by the integral nature of quark charges, the Han-Nambu approach came remarkably close to Quantum Chromodynamics (QCD) as understood today! These are the words Murray Gell-Mann had for this work [3]: “*If I had read Nambu’s paper, it would have set me ahead by several years... and, Nambu then invented what amounted to QCD, or the beginning of QCD in the 1966 paper...*”

All these issues were clarified to a great extent by Jogesh Pati and Abdus Salam in [28]. They were aiming at a unified picture of leptons and hadrons. By that time (1973) the theoretical foundations of the electroweak unification had been well established (see Sect. 2.7 of Chap. 2 of this book). Before going into more details of the Pati-Salam work, it is important to emphasize that electroweak unification had radically altered the picture of *flavour*: the old $SU(3)$ had to give way to a $SU(2)$ -doublet structure in which (u, d) had to be put into a doublet, and to handle the s -quark, a new quark, the c -quark had been postulated and eventually found, to make up the (s, c) doublet and there were to be more, the (b, t) doublet. Only the u, d quarks were light and isotopic spin, once viewed as fundamental to strong interactions was now viewed to be an *accidental symmetry*.

Pati and Salam incorporated this modern view of flavour by considering (u, d) and (c, s) quarks; they also enlarged the lepton families to include (e, ν_e, μ, ν_μ) . For technical reasons they consider a $SU(4)$ colour group whose $SU(3)$ subgroup they identify with the original Han-Nambu colour group. Next they *gauge* both the flavour $SU(4)$ and the colour $SU(4)$ by introducing 15-plets of gauge vector bosons for each of them. The flavour $SU(4)$ is spontaneously broken to generate *massive* vector bosons, except for the photon. Because of the inevitable colour-flavour mixing of the Han-Nambu scheme, it becomes inevitable to spontaneously break the colour group too, making the gluons massive too. Two immediate consequences were broken colour and massive gluons on the one side, and mixing of colour and electroweak gauge bosons on the other.

19.4.2 Fritzsche-Gell-Mann Approach

Gell-Mann’s original choice for quark dynamics was through the exchange of colour-singlet and electrically neutral vector gluons (see [29–31] for discussions of this earlier proposal). The vector interactions were chosen to preserve the observed *chiral symmetry* in low-energy hadronic interactions (see Chap. 18 for a discussion). There were many problems with this proposal. Chief among them being: (i) the gluons being colour singlet should be observable making them very likely massive, (ii) being massive they would not have been able to generate any infrared singularities conjectured by these authors to be the mechanism for suppressing colour non-singlet baryons and mesons from being observable (more on this issue later). Fritzsche in [1] states this another way: massive gluons can only generate short-range Yukawa potentials, not the kind that can deeply bind quarks inside hadrons, (iii) the colour-

singlet gluons would mix with vector mesons like ω , (iv) the desired deep attraction in colour-singlet channels would be absent, and (v) there would be a basic asymmetry between quarks which are coloured and gluons which are colourless. The last point is of course not particularly pressing as in QED too one has such an asymmetry. There are many other technical deficiencies like bad scaling behaviour, etc. L.B. Okun had also pointed out that when masses of u , d quarks are nearly equal, there would be an $SU(6)$ symmetry (of a different kind than what was once popular in non-relativistic quark model). The interested reader should refer to these papers.

So Fritzsche and Gell-Mann soon started considering the case where the gluons would be an octet of coloured vectors (electrically neutral). Even with an octet of gluons, they kept many options open for quark-gluon dynamics, with only one of them having the self-interactions of a Yang-Mills theory. In fact, they had even considered a dynamics incorporating the strings of Dual Resonance Models! [29].

An immediate advantage emerges over the Han-Nambu (Pati-Salam) dynamical models. Since colour and flavour are strictly decoupled in the Bardeen-Fritzsche-Gell-Mann colour scheme [24], the gauging of colour can be *exact* without the breaking in the flavour sector coming in the way! This means the gluons can be “massless”. Since they are coloured, principles requiring only colour-singlet states from being observable will prevent them from being directly observable. There will be no mixing between gluons and electroweak gauge bosons, nor with any vector mesons. Furthermore, the masslessness of the gluons along with the self-interactions among them has the potential of generating the type of infrared singularities alluded to above. The new dynamics also removes what Fritzsche calls an *annoying* asymmetry between quarks and gluons.

Again, vectors were chosen to reflect the observed chiral symmetry of the low-energy hadronic interactions (see Chap. 18 for a discussion). Many advantages of this colour octet gluon model of quark dynamics were described in detail by Fritzsche, Gell-Mann, and Leutwyler in their 1973 paper [31]. In addition to the advantages mentioned above, they pointed that one could get attraction in the colour-singlet channels, making the requirement that only colour-singlet states would appear as asymptotic states rather plausible. In support of such a picture they (as already mentioned) pointed out how such massless gluons could in principle generate the required infrared singularities. The colour symmetry being exact, there would be no mixing between colour-singlet and colour non-singlet states. They also pointed out that the observed scaling violations in deep inelastic scattering were closer to what this theory predicted. They make a number of very important technical observations about modifications to *Current Algebra*, etc. but we shall not go into them. Weinberg also discusses many advantages of QCD-like theories in [32]. He too raises the possibility of severe infrared singularities preventing colour non-singlet states from being observable.

While this so-called *infrared slavery* is rather appealing, it is also very hard to establish it theoretically. David Gross and Frank Wilczek in their detailed paper [33] on *Asymptotic Freedom* (more on it shortly) are rather blunt in pointing this out! To quote them, “..One clearly requires a dynamical explanation of such a miracle.” They try to link it to Asymptotic Freedom itself, but even that requires a dynamical

explanation which is lacking even to this day, nearly 50 years after the formulation of QCD. This is indeed the problem of *Quark and Gluon Confinement* (more on it later) and the final chapters of our book will have a lot of bearing on these issues. A particularly fruitful approach has been that of *Lattice Gauge Theories*, to be discussed in Chap. 20. This will also play a major role in the remainder of this book.

19.4.3 Observational Tests for Han-Nambu Model Revisited

Now we return to the apparent failure of the Han-Nambu integer charge quark model in accounting for the R-ratio. Pati and Salam [34], and independently Rajasekaran and Probir Roy [35] pointed out the amazing result that in Han-Nambu type of models, with intrinsic mixing of flavour and colour, spontaneous breakdown of both flavour and colour gauge invariances is also inevitable. Therefore, in addition to the weak gauge bosons becoming massive, gluons will also become massive. While all that is expected, what came as a surprise was the consequence that for momentum transfers $q^2 \gg m_g^2$ (m_g is the gluon mass), the contribution of colour to the electric charge is highly suppressed and deep inelastic scattering experiments of the type $e^+e^- \rightarrow \text{hadrons}$ will not easily distinguish Han-Nambu models from fractionally charged quark models.

More precisely, if $Q = Q_0 + Q_8$, where Q_0 are the fractional charges of the Gell-Mann-Zweig model, Q the charges in the Han-Nambu model, the effective charge seen at momentum transfer $Q_{eff}(q^2)$ is given by $Q_{eff}(q^2) = Q_0 + \frac{m_g^2}{m_g^2 - q^2} \cdot Q_8$. Therefore, the naive estimate for the R-ratio in Han-Nambu model is quite irrelevant! Rindani [36] discusses various possibilities for testing Han-Nambu models. It would also be worthwhile to investigate the *non-perturbative* aspects of this model through *Lattice Simulations*.

19.4.4 Deep Inelastic Scattering

We briefly mention some indirect evidences that pointed to electrically neutral objects contributing to the dynamics of quarks. This was that the momentum fraction carried by the electrically charged constituents of nucleons was only about 50%. These were revealed by the deep inelastic scattering experiments at the Stanford Linear Accelerator Centre around 1970 [37]. These same experiments also revealed the so-called Bjorken scaling, predicted by Bjorken in 1969 [38]. An interpretation of this scaling was given by the *Parton Model* of Bjorken and Feynman according to which the scatterers behaved like free pointlike objects [39,40]. Initially there was confusion whether these partons and quarks had anything to do with each other. Now it is believed they are the same. We shall see soon how the feature of *Asymptotic Freedom* of QCD provides natural explanations for all this.

19.5 Quantum Chromodynamics (QCD)

Now we finally come to a discussion of Quantum Chromodynamics (QCD) as finally formulated by Fritzsch, Gell-Mann, Bardeen and Leutwyler. Let us recapitulate the salient features: (i) the colour group is $SU(3)$ and is taken to be *exact*, (ii) the interactions between quarks are mediated by an $SU(3)$ octet of vector gauge bosons and (iii) colour and flavour act independently. The last point is a crucial feature of QCD. It means the colour gauge fields couple universally to all flavours of quarks. The so-called *Yang-Mills Construction* provides the simplest field theory description for such dynamics. C.N. Yang and R. Mills had shown how to construct such theories when the gauge group is $SU(2)$ [41]. Oskar Klein had given the same construction already in 1938 [42]. We begin by discussing how the essential features of this construction work for the simpler case of QED.

19.5.1 A QED Interlude

Consider the Lagrangean density for a free fermion field theory:

$$\mathcal{L} = \bar{\psi}(x) (i\gamma^\mu \partial_\mu - M)\psi(x) \quad (19.7)$$

This is obviously invariant under the *global* gauge transformations

$$\psi(x) \rightarrow \psi'(x) = e^{-iq\Lambda} \psi(x) \quad (19.8)$$

where Λ does not depend upon x . The infinitesimal form of this global gauge transformation is

$$\delta \psi(x) = -iq\Lambda \psi(x) \quad (19.9)$$

Now we turn to *gauging* this global invariance. What it means is how can we lift this to a local invariance, i.e. when $\Lambda(x)$ is position-dependent? Clearly additional fields are required and these are the *gauge fields*. The origin of the difficulty is that $\partial_\mu \psi(x)$ no longer transforms the same way as $\psi(x)$. The trick is to find a *gauge-covariant derivative* $D_\mu \psi(x)$ which indeed transforms the same way as $\psi(x)$. This is all standard text book material, so we just show the answer:

$$D_\mu = \partial_\mu + iq A_\mu(x) \quad (19.10)$$

where A_μ is the required vector gauge field with the transformation law

$$A'_\mu(x) = A_\mu(x) + \partial_\mu \Lambda(x) \quad (19.11)$$

which immediately leads to the gauge-invariant field strength

$$F_{\mu\nu} = \partial_\mu A_\nu(x) - \partial_\nu A_\mu(x) \quad F'_{\mu\nu} = F_{\mu\nu}(x) \quad (19.12)$$

culminating in the gauge-invariant interacting Lagrangean density

$$\mathcal{L}_{qed} = -\frac{1}{4} F_{\mu\nu}(x) F^{\mu\nu}(x) + \bar{\psi}(x)(i\gamma^\mu D_\mu - M)\psi(x). \quad (19.13)$$

19.5.2 QCD

Now we turn to a similar construction for QCD with the colour gauge group $SU(3)$. Again, all this is by now text book material, but has been covered here to make our book self-contained. Though what we want is a construction for $SU(3)$, we shall show how things work for the so-called *Simple* Lie groups; these are groups which are not of the direct-product type. The group $SU(3) \otimes SU(2)$ is not simple, but all $SU(N)$ groups are (for a good discussion of these classifications as well as a good introduction to Lie Groups see Appendix C of the book *Field Theory in Particle Physics* by B. de Wit and J. Smith [21]). We shall follow the notations of [17].

The Hermitean generators of the Lie Group (algebra) are denoted by $L_a, a = 1 \dots, n$. While physicists tend to use Hermitean generators, mathematicians prefer anti-Hermitean generators! Denoting the latter by T_a , the two can be related by $L_a = i T_a$. The dimension of the group $d(G) = n$. They obey the non-commutative (hence non-Abelian) Lie algebra

$$[L_a, L_b] = i f_{ab}^c L_c \quad (19.14)$$

f_{ab}^c are the *structure constants*, with the obvious antisymmetry $f_{ab}^c = -f_{ba}^c$. An important algebraic relation among generators is the *Jacobi Identity*:

$$[[L_a, L_b], L_c] + [[L_c, L_a], L_b] + [[L_b, L_c], L_a] = 0 \quad (19.15)$$

which immediately translates into the important identity among the structure constants

$$f_{ab}^d f_{dc}^e + f_{ca}^d f_{db}^e + f_{bc}^d f_{da}^e = 0 \quad (19.16)$$

As $SU(N)$ groups are *compact*, they admit unitary representations of finite dimensions. Denoting the dimension of representation r by $d(r)$, this means generators of representation r can be represented by $d(r) \times d(r)$ Hermitean matrices. The Jacobi identity among f_{bc}^a at once leads to the so-called adjoint representation with dimension equal to $d(G)$. To see that, introduce the matrices $L_{a,bc}^{ad} = -i f_{ab}^c$. It is then easy to rearrange Eq. (19.16) as

$$[L_a^{ad}, L_b^{ad}] = i f_{ab}^c L_c^{ad} \quad (19.17)$$

In other words, the matrices L_a^{ad} provide a matrix representation of dimension $d(r) = \dim(G)$, which is nothing but the *adjoint representation*. The hermiticity of the adjoint generators follows from the symmetry properties of the structure constants.

A very important concept is that of a *metric* for simple Lie algebras. It is defined by (also called the *Cartan-Killing metric* [21])

$$g_{ab} = \text{Tr}(L_a^{ad} L_b^{ad}) \quad (19.18)$$

This metric can be diagonalized to $g_{ab} = \delta_{ab}$. With this metric one can introduce $f_{abc} = f_{ab}^d g_{dc}$. The Jacobi identity can then be used to show that f_{abc} is *fully antisymmetric*. Because of the diagonal form of the metric it also follows that $f_{ab}^c = f_{abc}$ is also fully antisymmetric. It is also customary to choose normalizations such that, for the fundamental representation,

$$\text{Tr}_f(L_a L_b) = \frac{1}{2} \delta_{ab} \quad (19.19)$$

Now we have all the technical ingredients to develop the Lagrangean density for any non-abelian gauge theory based on a simple group; this can then be directly applied to the case of QCD where the relevant group is SU(3). Returning to the Lagrangean density for a free fermion field of Eq. (19.7) we take the fermion field $\psi(x)$ to now transform under some representation of the group; in the case of QCD, the fermions are the quarks transforming as the fundamental representation. Considering the unitary matrix $U = e^{-i\omega_a L_a}$, where L_a is the matrix for the representation of the fermion field, and ω_a are real parameters, the global gauge transformation

$$\psi'(x) = U \psi(x) \quad (19.20)$$

leaves the free fermion Lagrangean density invariant. Now the question is how to lift this invariance to a local one with $\omega_a(x)$ now depending on position. In other words, the local gauge transformation on fermionic fields, with $U(x) = e^{-i\omega_a(x)L_a}$, is

$$\psi'(x) = U(x) \psi(x) \quad (19.21)$$

Again, the constructions first given by Yang and Mills, and Oskar Klein are now text book material, so we just give the final results. This requires as many gauge vector bosons as $\dim(G)$, i.e. $A_a^\mu(x)$, $a = 1 \dots, \dim(G)$. Introducing the matrix-valued gauge field $\mathbf{A}^\mu(x) = A_a^\mu(x) L_a$ where L_a is again in the same representation as the fermions, the gauge-covariant derivative takes the form

$$D_\mu = \partial_\mu + ig \mathbf{A}_\mu(x) \quad (19.22)$$

with the desired transformation law

$$(D_\mu \psi)' = U(x) (D_\mu \psi) \quad (19.23)$$

resulting in the transformation law for gauge fields themselves:

$$\mathbf{A}'_\mu(x) = -\frac{i}{g} U(x) \partial_\mu U^\dagger(x) + U(x) \mathbf{A}_\mu(x) U^\dagger(x) \quad (19.24)$$

The case of QED could also have been obtained by choosing $U(x)$ to be the phase $U(x) = e^{-iq\Lambda(x)}$. A very important difference is to be noticed over the abelian case: in addition to the inhomogeneous first term which is present in the abelian case too, there is a homogeneous part to the transformation law too. This will result in crucial differences including self-interactions among the gauge bosons with far-reaching dynamical consequences.

What about the analog of $F_{\mu\nu}$? A naive generalization to $\mathbf{F}_{\mu\nu}(x) = \partial_\mu \mathbf{A}_\nu(x) - \partial_\nu \mathbf{A}_\mu(x)$ has no nice transformation properties. Instead, what transforms nicely is the combination

$$\mathbf{F}_{\mu\nu}(x) = \partial_\mu \mathbf{A}_\nu(x) - \partial_\nu \mathbf{A}_\mu(x) + ig[\mathbf{A}_\mu(x), \mathbf{A}_\nu(x)] \quad (19.25)$$

with the beautiful transformation law

$$\mathbf{F}'_{\mu\nu}(x) = U(x) \mathbf{F}_{\mu\nu}(x) U^\dagger(x) \quad (19.26)$$

As a consequence, one arrives at the important gauge-invariant quantity

$$\text{Tr}(\mathbf{F}_{\mu\nu} \mathbf{F}^{\mu\nu}) \quad (19.27)$$

In terms of the component fields $A_a^\mu(x)$, the components of the field-strength tensor take the form

$$F_a^{\mu\nu}(x) = \partial^\mu A_a^\nu(x) - \partial^\nu A_a^\mu(x) - g f_{abc} A_b^\mu(x) A_c^\nu(x) \quad (19.28)$$

Finally, the gauge-invariant Lagrangean density for non-abelian gauge theories is

$$\mathcal{L} = -\frac{1}{4} F_a^{\mu\nu} F_{a\mu\nu} + \bar{\psi}[i\gamma^\mu(\partial_\mu + ig A_\mu^a L_a) - M]\psi \quad (19.29)$$

The all important self-interactions of the gauge field are contained in the F^2 -terms. As the coupling g and A always come in a product (this is true for QED too) some people absorb them together and then the entire g -dependence occurs as a scale factor $\frac{1}{g^2}$ in $\mathcal{L}$. We shall not pursue this point of view anymore. From a *Renormalization Group* point of view, it has some advantages.

Coming to QCD, the group in question is SU(3), the fermions are quarks in the fundamental, i.e. **3** representation of SU(3) and L_a are the generators of SU(3) in the fundamental representation.

There are so many aspects of QCD, all very interesting and exciting, but we shall limit our discussions to only those that are central to the narratives of this book. Therefore we shall limit ourselves to (i) Asymptotic Freedom, (ii) the static quark-antiquark potential in QCD, (iii) the problem of quark confinement and (iv) Chiral Symmetry in QCD. We shall also be concerned with only the pure-gauge-field aspects.

19.6 Asymptotic Freedom and Perturbation Theory

19.6.1 Sliding Scales and Running Couplings

In this section, we take up several issues surrounding the validity of perturbation techniques in Quantum Field Theories, a technique that had impressive achievements in QED. This, among other things, is contingent on the “coupling constants” being small. But, as Weinberg has pointed out in his introduction to his Chap. 18 on Renormalization Group, smallness of coupling does not suffice as often it is accompanied by the “large logarithms” or in massless theories like QCD one encounters logarithms that become too large at very high and very low energies [43]. In QED the coupling constant is characterized by the well-known *fine-structure constant* $\alpha = \frac{e^2}{\hbar c} \simeq \frac{1}{137}$, and is just a number. A formal way to understand is that in QED, after regularizing the theory with a cut-off, the process of *renormalization* involves defining the physical charge (or coupling constant) as the $q^2 = 0$ value of the photon-charge particle vertex function. The numerical value of this *renormalized charge* is then determined observationally. But a moment’s reflection makes it clear that the renormalized charge could equally well have been defined by evaluating the regularized vertex function at any value of $q^2 = \mu^2$! Then too it’s value would subsequently be determined through experiments. The electric charge e_μ so defined would naturally depend on the value of the “scale” μ . Likewise, in $\lambda \phi^4$ scalar field theories, the value of the renormalized λ_R can be defined to be the value of the four-point amplitude at any $s = t = u = \mu^2$. In the particular example of QED, the point $q^2 = 0$ is *scale invariant*.

In the case of QCD with massless self-interacting gluons, the infrared singularities of the theory do not even permit the luxury of choosing such a scale-invariant definition of the coupling constant, and one necessarily has to define it at an arbitrary scale. The arbitrariness in the scale is itself referred to as the *sliding scale*. In terms of the renormalization of quantum field theories, the arbitrariness of scale is a reflection of the freedom to perform *finite* renormalizations. Now it is clear that even QED can be thought in such a general way and that the special value of $\simeq \frac{1}{137}$ only reflects a special value of the scale.

19.6.2 The Callan-Symanzik Equation and the Beta Function

Once one admits this additional freedom in defining couplings, the natural next question is how to guarantee that “physics” (for example, the S-matrix) is independent of such choices, as it indeed ought to be. Addressing this important issue was the objective of the so-called *Renormalization Group* (a name first coined by Stueckelberg and Peterman [44], which Weinberg [43] rightly calls *unfortunate* as there can be no group underlying the transformations! But the name is doing well even after 70 years!). By couplings, we mean here the *renormalized couplings* of QFT.

In fact, not just couplings but all Green’s functions will also depend on the sliding scale, and together they must conspire to leave the “physics”, more precisely the

S-matrix, unchanged. Stueckelberg and Peterman were the first to investigate this in 1953. Soon afterwards, Murray Gell-Mann and Francis Low [45] gave their famous renormalization group analysis which was based on the photon two-point function. The reader is referred to [17, 43, 46, 47] for a good discussion. I would strongly recommend the book by de Wit and Smith [21] for all aspects of the renormalization group as they have detailed discussions with very explicit derivations. This is so for various schemes, in particular, the minimal subtraction scheme, the running coupling constants, etc. A strong motivation for Gell-Mann and Low was also as a way to improve perturbation theory. The reader is recommended to the book by Bjorken and Drell [47] for a highly illuminating discussion. Bogoliubov and Shirkov developed these ideas further [48]. Almost a decade later, these ideas were revived by Curtis Callan [49, 50] and Kurt Symanzik [51]. These resulted in the now famous *Callan-Symanzik equations* as the conditions for keeping physics unchanged as one changes the scale, which is also physical. A parallel development was that of Kenneth Wilson [52, 53]. In addition to the sources mentioned above, we refer the reader further to Sivakumar's book [54] for a lucid description and derivation of these equations. We shall not go into the details of this equation. We shall, however, write it down for making the chapter self-contained. To convey the essentials, it suffices to look at it for one of the simplest field theories, viz. $g\phi^4$ theory with coupling g and mass m (we have followed the treatments given in [21, 54]):

$$\left\{ \mu \frac{\partial}{\partial \mu} + \beta(g, \mu) \frac{\partial}{\partial g} + \gamma_m m \frac{\partial}{\partial m} + \frac{n}{2} \gamma \right\} \Gamma^{(n)}(p_i; g, m, \mu) = 0 \quad (19.30)$$

Compared to [21] we have allowed for scale dependence of masses also. The coefficient functions are defined according to

$$\beta(g, \mu) = \mu \frac{\partial g}{\partial \mu} \quad \gamma_m = \mu \frac{\partial \ln m}{\partial \mu} \quad \gamma = \mu \frac{\partial \ln Z}{\partial \mu} \quad (19.31)$$

with Z being the wavefunction renormalization. For QED, there will be two wavefunction renormalizations, one for the photon field and another for the electron field. Likewise in QCD for quarks and gluons.

It should be noted that the coefficient functions do not depend on the choice of the Green function $\Gamma^{(n)}$. That suggests using the simplest Green function, namely, the propagation function. In fact, that was exactly the starting point of the original RG work by Gell-Mann and Low [47]! The meaning of the Callan-Symanzik function becomes immediately transparent then. As the static quark-antiquark potential of QCD will be among our big focus, and as that is essentially the gluon propagator (exact), we will elaborate on this aspect. We shall first consider the simpler situation of the exact renormalized photon propagator in QED, which can be expressed as (see Sect. 19.15 of [47]):

$$e_R^2 \tilde{D}'_F(q)_{\mu\nu} = -\frac{g_{\mu\nu}}{q^2} d(q^2, 0, e_R^2) + \text{gauge terms} \quad (19.32)$$

In the above, the renormalized coupling e_R has been defined at the scale $\mu = 0$ and is to be identified with $\frac{e_R^2}{4\pi} \simeq \frac{1}{137}$. More precisely,

$$d(0, 0, e_R^2) = e_R^2 \quad (19.33)$$

Now suppose we define the renormalized charge at some other sliding scale μ and let e_μ denote the corresponding renormalized charge. Then the above equation would change to

$$d(0, \mu^2, e_\mu^2) = e_\mu^2 \quad (19.34)$$

The potential between charges is observable and this cannot depend upon the arbitrary choice of the sliding scale. This is expressed by the equation

$$d(q^2, \mu_1^2, e_{\mu_1}^2) = d(q^2, \mu_2^2, e_{\mu_2}^2) \quad (19.35)$$

This is the essential content of the Callan-Symanzik equation. In particular, the couplings at sliding scales have no observable meaning by themselves. The precise statement of the Callan-Symanzik equation for $d(q^2, \mu, g)$ is, however,

$$\left(\mu \frac{\partial}{\partial \mu} + \beta(g) \frac{\partial}{\partial g} + \gamma(g) \right) d(p_i, \mu, g) = 0 \quad (19.36)$$

We have used $n = 2$ as appropriate for a two-point function and have dropped the mass-related terms. The latter can be justified when either there are no masses or when the scales are much larger than the masses. We shall return to these subtleties later.

19.6.3 The Running Couplings

Now we come to a different, but somewhat related, concept of a *Running Coupling*. Though it emerges from pure dimensional analysis, it will prove to be highly useful. Let us again start with the effective coupling $d(q^2, \mu^2, g_\mu^2)$. Though we motivated it based on QED, its validity is more general. Let us keep the scale μ *fixed*, but ask what happens if we keep scaling only q^2 . Let us for the moment consider only increasing q^2 . Since the coupling $d(q^2, \mu^2, g_\mu^2)$ is *dimensionless* (the arguments are easily generalized to renormalized Green functions that are dimensionful; see [21] Sect.9.3), it can only depend on the ratio $\frac{q^2}{\mu^2}$. Recall the elementary differential identity satisfied by any function of the type $f(\frac{x}{y})$:

$$\left(x \frac{\partial}{\partial x} + y \frac{\partial}{\partial y} \right) f(x/y) = 0 \quad (19.37)$$

Consequently, $d(p_i, \mu, g)$ satisfies (we have traded q^2 , without loss of generality, for the relevant momenta p_i)

$$\left(\mu \frac{\partial}{\partial \mu} + \frac{\partial}{\partial t} \right) d(e^t p_i, \mu, g) = 0 \quad (19.38)$$

It should be noted that t is dimensionless. On combining this with the Callan-Symanzik equation for d of Eq. (19.36), one gets

$$\left(\frac{\partial}{\partial t} - \beta(g) \frac{\partial}{\partial g} - \gamma(g) \right) d(e^t p_i, \mu, g) = 0 \quad (19.39)$$

The prime motivation for introducing the concept of *running coupling* comes from this equation. This involves introducing the function $g(t)$ required to satisfy

$$\frac{d}{dt} g(t) = \beta(g(t)) \quad g(0) = g \quad (19.40)$$

It is then an easy exercise to show (see [21]) that the function $F(t)$ defined by

$$F(t) \equiv e^{\int_0^t d\tau \gamma(g(\tau))} d(e^{-t} p_i, \mu, g) \quad (19.41)$$

is actually independent of t . Equivalently,

$$d(e^t p_i, \mu, g) = e^{\int_0^t d\tau \gamma(g(\tau))} d(p_i, \mu, g(t)) \quad (19.42)$$

The exponential multiplicative factor reflects the fact that though the coupling constants have mass dimension zero (also called engineering dimension), their anomalous dimensions need not vanish. Though we have shown how this works for the two-point function, it can be established with equal ease for arbitrary Green functions [21]. Since the definition of the running coupling as given by Eq. (19.40) depends only on the beta function which in turn is independent of the process under consideration, it follows that the same $g(t)$ governs the scaling behaviour of all processes. This brings out the significance of the running coupling; Green functions at a scaled momentum $e^t p_i$ can be obtained by merely changing the coupling g to $g(t)$, at p_i !

19.6.4 The Beta Function

The solution of the Callan-Symanzik equations requires the knowledge of the so-called *coefficient functions* β, γ, γ_m , etc. These are, respectively, called the *Beta Function* (not to be confused with the Euler Beta function!), the *Anomalous Dimension*, and the *mass dimensions*. The last arises from the fact that even the masses in a quantum field theory depend on the scale. We shall be mostly concerned with only

the β -function in what follows and give its definition as follows from the CG-eqn [43,54]:

$$\beta(g_\mu) \equiv \mu \frac{\partial g_\mu}{\partial \mu} \quad (19.43)$$

where g_μ is the coupling at scale μ . Though we have treated only one coupling, this can be easily generalized to many couplings. It should be noted that Eq. (18.2.9) of [43] is not the CG-eqn as Weinberg states and is in fact the definition of the β -function. Knowing this function, one can integrate it to get the explicit dependence of the couplings on the sliding scale. Techniques for calculating this are by now standard.

A feature of importance, though seemingly elementary but a source of many confusions, should be pointed out at this stage. Scaling μ does not change the equation. It is actually a first-order differential eqn. in $\log \mu$. But $\log \mu$ does not make any dimensional sense. Therefore this equation should be understood as being in $\log \frac{\mu}{\mu_0}$ where μ_0 is an arbitrarily specified scale. This is the price for associating a scale dependence with a dimensionless quantity such as the gauge coupling in $D = 4$. In cases like QED in $D = 3$, where the gauge coupling is itself dimensional, the discussions take totally different forms.

Few more important remarks are in order here. Though one way of defining renormalized couplings by the values of some physical processes, for example, scattering, at some freely chosen scale, is a possibility, it is not the easiest way for determining the β -functions. In perturbative QFT's that are renormalizable, the most practical way is through the various renormalization factors like charge renormalization, wave function renormalization, etc. This is a two-step process in that the divergences of the field theory have to be regularized first and then the observable parameters like mass, charge etc. have to be renormalized. There are many ways of doing either of these. All of them can in principle give rise to different β -functions. Different schemes are related to each other by *finite* renormalizations. In gauge theories like QED and QCD, there are additional complications, not surprisingly! While the physical processes are gauge-invariant, and couplings defined through them are also gauge-invariant, leading to gauge-invariant β -functions, the methods using renormalization factors are not always so, and, in principle, the β -functions can be gauge dependent.

Gerard t'Hooft and Martinus Veltman [55] in 1972 introduced a novel regularization scheme called *dimensional regularization* wherein the space-time dimensions are taken to be $D = 4 - \epsilon$ with ϵ small and positive. This was a gauge-invariant regularization scheme so the β -functions will also be gauge independent, as was shown by Caswell and Wilczek [56]. In his seminal work on *Dimensional Regularization and Renormalization Group* in 1973, t'Hooft had introduced the *Minimal Subtraction*(MS) scheme [57], explicitly linked to dimensional regularization, whereby only the $\frac{1}{\epsilon}$ parts of the renormalization factors were absorbed to define the renormalized couplings. This also came to be known as the *mass independent* scheme as the β -functions would not depend on any masses of the theory [57]. In that paper the reasons for mass independence were buried in a lot of technical details. Satish Joglekar reanalysed the mass-independence issue in Yukawa-like theories to make the issues

more transparent [58]. A closely related scheme is the $\overline{\text{MS}}$ scheme, also called the *Modified Minimal Subtraction Scheme*, where not just the divergent $\frac{1}{\epsilon}$ piece, but also some constants that always accompany them like the Euler-Mascheroni constant γ_E are also absorbed into the definition of the renormalized quantities. We again refer the reader to [21] for a lucid treatment and worked out examples.

The minimal subtraction schemes MS and $\overline{\text{MS}}$ have many great advantages. The beta functions and anomalous dimensions being mass independent, the Callan-Symanzik equations can be integrated exactly. They suffer from disadvantages too. The very mass independence obscures mass effects in physical observables. For example, threshold behaviours associated with mass hierarchies are not at all transparent. Nevertheless, it has become one of the most popular schemes. This will be explicitly illustrated in the case of QED shortly.

Now we summarize the broad features of β -functions (perturbative) as far as these issues are concerned:

- The first term is both gauge independent and scheme independent in *all* schemes.
- The second term is also scheme independent.
- In the MS and $\overline{\text{MS}}$ schemes, the entire β -function is gauge independent.

In a very important but unpublished work in 1972, Symanzik pointed out the possibility of field theories whose couplings decrease at high energies. This observation hinges on the properties of the β -function, and can be understood on rather general terms. Before proceeding further, we pause to point out that QCD possesses this property with deep ramifications, and it is called *Asymptotic Freedom*.

There are essentially three possibilities (see [43,54] for details): (i) $\beta(g) = 0$, these are scale-invariant theories. Examples of this kind are some *Supersymmetric Theories*, (ii) $\beta(g) < 0$: in the vicinity of couplings where this happens, the coupling decreases with increase of scale. If this happens at small enough g_μ , one may expect the coupling to vanish asymptotically with energy. If this happens at finite g_μ , it is hard to say what happens, (iii) $\beta(g) > 0$, in this case the coupling increases with increasing scale. QED is of this kind, and such theories are pathological.

As we have just seen, the same beta function serves dual purposes as reflected in Eqs. (19.40), (19.43) with entirely different meanings. Both these equations are exactly of the same form and this can be a source of confusion between the meaning of their solutions. Now we shall integrate both these types of equations to bring out their significances. We first take up the version related to the dependence of couplings on the sliding scale.

Let us examine the β -functions for QED and QCD in leading order.

19.6.5 RG in QED

For QED (see [43]),

$$\beta_{qed}(e) = \frac{e^3}{12\pi^2} > 0 \quad (19.44)$$

Even without explicitly integrating Eq. (19.43) it is clear that the coupling e_μ increases with increasing μ . On introducing $\bar{t} = \log \frac{\mu}{\mu_0}$, with μ_0 an arbitrary unit, (we have distinguished this from t introduced while discussing the running coupling with a) Eq. (19.43) is recast as:

$$\frac{d e(\bar{t})}{d \bar{t}} = \frac{e(\bar{t})^3}{12\pi^2} \quad (19.45)$$

The solution with the initial condition $e(\bar{t} = 0) = e_0$ is

$$e^2(\bar{t}) = \frac{e_0^2}{1 - \frac{\bar{t}}{6\pi^2} \cdot e_0^2} \quad (19.46)$$

It is clear that as $\bar{t}$ increases, $e(\bar{t})$ increases. In fact, it blows up at a scale $\bar{t} = \frac{6\pi^2}{e_0^2}$. This is called the *Landau Singularity*. But by the time the singularity is reached, perturbation theory will have broken down and the expression for $\beta(e)$ used totally unreliable.

In Eq. (19.46), e_0 has nothing to do what one may call e_R of QED related to the fine-structure constant $\alpha = \frac{e_R^2}{4\pi} \simeq \frac{1}{137}$. That is the renormalized charge at the scale-invariant point $\mu = 0$. It is quite clear from the solution above that $\mu = 0$ can never be accessed as $\bar{t} = -\infty$ there. In fact the renormalized charge though it keeps decreasing with decreasing $\bar{t}$ does not reach any fixed value. It instead keeps decreasing!

The resolution of this lies in the fact that the beta function of QED given above is *mass independent*. It is strictly valid when there are no mass scales or when μ is much larger than all the mass scales. Weinberg in Sect. 18.2 of [43] has explicitly clarified this situation. In perturbative QED to leading order, with the electron mass occurring explicitly, the relationship between e_μ and e_R is:

$$\frac{e_R}{e_\mu} = 1 - \frac{e_R^2}{4\pi} \int_0^1 dx x(1-x) \ln \left[1 + \frac{\mu^2 x(1-x)}{m_e^2} \right] + \dots \quad (19.47)$$

When evaluated for μ much larger than m_e , but not so large as to make the perturbative correction comparable to unity, this indeed reproduces the beta function of Eq. (19.45), but otherwise it exists for $\mu = 0$. This should sensitize the reader as to the caution to be exercised when using beta functions.

Let us now take a look at the evolution of the running coupling constant in QED $e(t)$. As the functional form of the two differential equations are the same, we need not solve for $e(t)$ again, but simply take over the solution of Eq. (19.46) with t replacing $\bar{t}$ and some reinterpretations. Recall that now we are discussing how the coupling changes with momentum scale while the sliding scale is fixed, say, at μ :

$$e^2(t) = \frac{e_\mu^2}{1 - \frac{t}{6\pi^2} \cdot e_\mu^2} \quad (19.48)$$

While the $\bar{t} = 0$ earlier referred to some arbitrary unit μ_0 , $t = 0$ now refers to the momentum scale being the same as the sliding scale μ , which can however be chosen arbitrarily. Consequently t can be identified with $\frac{1}{2} \ln \frac{q^2}{\mu^2}$. Therefore QED has the feature that the running coupling increases with increase of scale. This feature was well known for a long time as can be seen from the extensive discussions in [47]. If continued this will of course hit the Landau singularity. But long before such a high momentum scale is reached, QED itself can get modified as indeed happens with the electroweak unification. But similar line of thinking can be applied to the couplings of the electroweak theory. We shall return to this after we have analysed the beta function for QCD.

Though we integrated this equation with the initial condition at $t = 0$, it is not necessary to do so. The initial condition could have been at some arbitrary $t_0 = \ln \frac{\mu_0}{\mu}$ with μ_0 quite distinct from the sliding scale μ . In that case, the expression for $e(t)$ would have taken the form:

$$e^2(q^2) = \frac{e_{\mu_0}^2}{1 - \frac{\ln \frac{q^2}{\mu_0^2}}{12\pi^2} \cdot e_{\mu_0}^2} \quad (19.49)$$

Now all reference to the sliding scale has disappeared.

19.6.6 Improving Perturbation Theory

One of the prime motivations of Gell-Mann and Low in their pioneering work on RG [45] was actually to improve perturbation theory. How this (almost miraculous result) can be carried out in practice is spelt out in detail in the book by Bjorken and Drell (see Eqs. (19.151)–(19.157) of [47]). They discuss how, starting with the effective charge calculated to fourth order, one can calculate the *leading logarithmic behaviour* to sixth order, which happens to be $\ln \frac{q^2}{\mu^2}$. However, the sub-leading log behaviour, in this case $\ln \frac{q^2}{\mu^2}$ at sixth order cannot be determined this way. The reader is also referred to Weinberg [43] (Sect. 18.8) for further elucidation. As Weinberg remarks, none of this would have been easy to get without the RG methods. In all this, great care must however be exercised to make sure terms are not neglected which are of the same order as those retained.

19.6.7 RG in QCD

Now we turn to what happens in the case of QCD. The first calculation showing that $\beta(g) < 0$ was done by Gerard t'Hooft in 1972. He did not publish his results but had announced that $\beta(g) < 0$ for QCD at a conference on Gauge theories at Marseilles in 1972 [59]. D.J. Gross and Frank Wilczek, and H.D. Polizer independently discovered Asymptotic Freedom in 1973 [60, 61]. Gross and Wilczek elaborated the many facets

of Asymptotic Freedom in two long papers [62,63]. We start with the leading order (in perturbation theory) β -function for a SU(3) non-abelian gauge theory with n_f number of quark flavours and n_s number of coloured scalar “quarks”:

$$\beta(g_\mu) \equiv \mu \frac{\partial g_\mu}{\partial \mu} = - \left(11 - \frac{n_s}{6} - \frac{2n_f}{3} \right) \frac{g_\mu^3}{16\pi^2} \quad (19.50)$$

It should be noted that the non-gauge matter fields work oppositely to the gauge fields irrespective of whether they are fermions or bosons. This result, being leading order, holds for all schemes and is also gauge independent.

We shall look at several ways of integrating the beta function to get the running coupling for QCD. We shall not do the similar exercise for the sliding scale case. Our motivation for looking at the problem in different ways is largely motivated by the desire to clarify the meaning of the so-called QCD scale Λ . Introducing $b_0 = 11 - \frac{2n_f}{3}$ (in QCD there are no coloured scalars so we have set $n_s = 0$), which is *positive* even when all the known quark flavours are taken into account at $n_f = 6$, the running coupling differential equation becomes

$$\frac{d g(t)}{dt} = - \frac{b_0}{16\pi^2} g(t)^3 \quad (19.51)$$

It is the same form as in the QED case though with a different coefficient and, most importantly, with opposite sign! Let μ once again be the sliding scale. Integrating this from $t_0 = \ln \frac{\Lambda}{\mu}$, where we have deliberately renamed the arbitrary initial momentum scale as Λ instead of μ_0 to sharpen the discussion of the so-called *QCD* Λ , one gets

$$g^2(q^2) = \frac{g_0^2}{1 + g_0^2 \frac{b_0}{16\pi^2} \ln \frac{q^2}{\Lambda^2}} \quad (19.52)$$

We again emphasize that Λ introduced here is an initial momentum scale that is otherwise totally arbitrary.

In many ways the discussion of the asymptotic behaviour of $g(q^2)$ is on much better ground in QCD than in QED. Because of the negative sign of the perturbative beta function, equivalently the positivity of b_0 , the denominator never vanishes. So q^2 can be taken to arbitrarily large values without worrying about Landau-like singularities. Even without integrating the beta function, its negative sign immediately indicates a *decrease* in coupling as one goes to higher and higher momentum scales. Therefore, at sufficiently large momenta the coupling must be small enough to justify perturbation theory! This is the essence of *Asymptotic Freedom* (AF), which we shall elaborate shortly.

Now, we have integrated the leading order beta function as if it were an exact expression mathematically. But it is only meaningful as the first term in a perturbative expansion. For self-consistency, the solution must also be interpreted in the same

spirit, and only its leading order should be retained. Let us envisage working at a large enough q^2 such that $\ln \frac{q^2}{\Lambda^2} \gg \frac{16\pi^2}{b_0 g_0^2}$, so one gets

$$\alpha_s(q^2) \equiv \frac{g(q^2)^2}{4\pi} = \frac{4\pi}{b_0} \frac{1}{\ln \frac{q^2}{\Lambda^2}} \quad (19.53)$$

This is the familiar leading order asymptotic freedom formula.

We have repeatedly stressed that in the above derivation Λ is only a reflection of the initial conditions for iterating the beta function and arbitrary as such. Nevertheless, in the early days there were serious attempts to determine Λ from the above formula, and interpret it as some sort of physical scale associated with QCD. To clarify the subtleties involved, let us first examine the Eq. (19.51). In this, if we alter the initial conditions, both g_0 , Λ change in such a way as to leave $g(t)$ *unchanged*! Yet, Eq. (19.53), which is the leading order approximation, if literally used as a mathematical formula, would give the superficial impression that as Λ changes, $g(t)$ would change. In particular, this gives the erroneous impression that an experimental determination of $g(t)$ would yield Λ . The fallacy in this is not hard to see, and was first pointed out by Marjan Bace [64]; if we change Λ to another Λ' , though as a formula $g(t)$ would change too, not so if we keep in mind that self-consistency would demand that we keep only the leading order change, which of course vanishes! Thus any experimental determinations of Λ would only make sense if beta functions at higher orders are used.

Another interpretation of Λ is that offered by Weinberg (see the remarks after Eq. (18.7.7) of [43]) as a constant of integration. Since this view applies equally well when even beta functions at higher order are used, we undertake a generic analysis. An indefinite integration would yield

$$\int \frac{dg(t)}{\beta(g(t))} = t + C \quad (19.54)$$

with C being the constant of integration and the dimensionless variable $t = \ln \frac{q}{\mu}$ and μ the sliding scale. Now choosing this constant of integration to be parametrized as

$$C = -\ln \frac{\Lambda}{\mu} \quad (19.55)$$

brings the solutions to the forms obtained earlier. Of course, this does not in any way change the interpretation of what Λ is. It is still arbitrary without any dynamical interpretation as a *physical scale* of QCD like proton or pion mass or a scattering length. Even as a constant of integration it needs an initial condition for its determination and all the earlier remarks apply.

This “appearance” of a dimensionful parameter Λ in an otherwise scale-less theory like QCD (without fermions or with massless fermions, for example) has led to fancy concepts like *Dimensional Transmutation*. But the origin of this must be fairly clear now. In particular, the recognition that this scale has no dynamical significance as

such. In the context of sliding scales, it is a reflection of the fact that only comparison of couplings at two sliding scales has a meaning. In the context of running couplings, it is a reflection of the fact that the sliding scale can always be traded off for another scale.

19.6.8 Higher Order Beta Functions

Soon after t'Hooft's unpublished calculation of the leading order QCD beta function [59], the works of Gross and Wilczek [60] and of Politzer [61], Caswell [65], Jones [66], and Egorian and Tarasov [67] calculated the two-loop beta functions for SU(N) non-abelian gauge theories with fermionic matter. It took another 6 years for Tarasov et al. to calculate the three-loop QCD beta function [68]. The complexity of the calculations rapidly grows with the number of loops. J.A.M. Vermaseren pioneered powerful symbolic manipulation software like FORM [69] that had a dramatic impact on speeding up these calculations. Thirteen years after the Tarasov et al. three-loop calculation, Larin and Vermaseren confirmed the results using FORM in 1993 [70]. But it took just 4 more years for van Ritbergen, Vermaseren and Larin to carry out the formidable four-loop beta function calculation [71]. With calculations becoming so extraordinarily complex, it becomes very important to have such calculations verified by as many as possible. So it is gratifying that Czakon indeed independently verified the four-loop calculations [72] almost a decade later.

For the purposes of this book which seeks to establish an effective string theory description of the flux tubes of QCD, in the spirit of effective field theory descriptions elaborated in Chap. 18, the short-distance descriptions of which are provided by the asymptotically free end of QCD, it suffices to restrict attention to just the two-loop expressions. This is because the flux tubes will be probed via the static quark-antiquark potential, which is essentially the gluon two-point function (which is also essentially the effective charge). For $D = 4$ this potential becomes *infrared divergent* at three loops and beyond [73]. For $D = 3$, the infrared divergence occurs at two-loop level itself [74]. The two-loop beta function for $D = 4$ QCD [65] is:

$$\beta(g) = -b_0 \frac{g^3}{16\pi^2} - b_1 \frac{g^5}{(16\pi^2)^2} \dots \quad (19.56)$$

with the coefficients b_0, b_1 for SU(N) gauge theory with n_f number of quarks given by

$$\begin{aligned} b_0 &= \frac{11}{3} C_2(G) - \frac{4}{3} T_f n_f \\ b_1 &= \frac{34}{3} C_2(G)^2 - \frac{20}{3} C_2(G) T_f n_f - 4 C_2(R) T_f n_f \end{aligned} \quad (19.57)$$

where $C_2(G) = N$, $T_f = \frac{1}{2}$, $C_2(R) = \frac{N^2-1}{2N}$ for $SU(N)$ (C_2 are the so-called *Quadratic Casimirs*). For $SU(3)$, relevant for QCD, these are:

$$b_0 = \frac{11}{3} - \frac{2n_f}{3} \quad b_1 = 102 - \frac{38n_f}{3} \quad (19.58)$$

So even for $n_f = 6$, both b_0, b_1 are positive and two-loop corrections further enhance the tendency towards asymptotic freedom.

19.6.9 More on Asymptotic Freedom

The essence of asymptotic freedom in QCD, as we have seen so far, is that as the momentum scale goes up the running coupling decreases. This means the quarks become essentially free during large momentum transfer scatterings. But it is important to bear in mind that the running coupling remains finite, though very small, asymptotically. In fact, the departure from completely free behaviour is logarithmic as can be seen from Eq. (19.53). Now let us recall the experimental and theoretical situation regarding the so-called *deep inelastic scattering* around 1970 [37–40]. We have already briefly alluded to these before, but now we wish to make the discussion more quantitative.

The experiment involves scattering of very high-energy electrons off protons. The scattering is characterized by the energy E of the electron in the “lab” (proton rest frame) frame, energy E' of the scattered electron in the same frame and the angle through which the electron is scattered, also in the same frame. The electron’s energy loss is $\nu = E - E'$, and the momentum transfer is denoted by Q^2 . The dimensionless variables $x = \frac{Q^2}{2M\nu}$, $y = \frac{\nu}{E}$ play important roles. In the very high-energy regime, the initial proton is converted into many hadrons H which are not individually monitored. Instead, the total cross section is measured of the inelastic processes. This *inclusive* cross section where individual hadrons in the final state are not tagged can be related to the amplitude for virtual Compton scattering of protons. In contrast, exclusive processes where some of the final state hadrons are kept track of are much harder to describe. The Compton scattering is characterized by two invariant *Structure Functions* W_1, W_2 both functions of (ν, Q^2) . The reader is recommended to consult Weinberg’s book [43] (Sect. 20.6), the book by de Wit and Smith [21] (Sect. 6.54), and [46] (Sect. 13.4) for more details.

The big surprise from the Stanford experiments was that $\nu W_2(\nu, Q^2)$ in the regime where both ν, Q^2 are large seemed to depend only on x , with a mild dependence on Q^2 [75]. This came to be known as *Scaling*, more precisely *Bjorken Scaling* as Bjorken had been anticipating such scaling behaviour on the basis of Current Algebra [38]. Feynman had a more *intuitive* (in the words of Weinberg [43]) explanation according to which hadrons, in so far as deep inelastic scatterings are concerned, behave as if composed of non-interacting *partons* [39, 40]. Now called the Bjorken-Feynman Parton model, it introduced such notions as *Parton Distribution Functions* in terms of which many phenomena could be accounted for.

The surprising part is the almost non-interacting behaviour of the quark-partons. Why was the current algebra approach of Bjorken hinting at that? Possibly because the spirit of current algebra was to work out the algebra of currents based on a free quark field theory and to use general principles like Lorentz invariance to infer matrix elements of currents between hadronic states, etc. It also turns out that the relevant parts of current-commutators are those at short distances. Wilson developed his powerful techniques of *Operator Product Expansions* [76,77] which clarified many of these issues. This picture of non-interacting quark-partons was called the *Naive Quark-Parton Model*. Very soon, by 1970 or so, more accurate data revealed the presence of mild *scaling violations* [37,78].

The discovery of Asymptotic Freedom in QCD at once provided a natural basis for understanding the deep inelastic data. The decrease of the running coupling at high momentum transfers explained the nearly free behaviour of the quark-partons. But QCD also predicted a weak, logarithmically falling off interaction. This meant that the quark-partons were not exactly free and this could be a potential explanation for the observed scaling violations. As to the exact degree and nature of scaling violations, a lot of hard work based on QCD was necessary. Gross and Wilczek gave a very detailed analysis of the deeper implications of Asymptotic Freedom in two foundational papers [60,63]. Georgi and Politzer [79] also demonstrated many important aspects of AF.

Most salient among these are that QCD predicts a logarithmic scaling violation. Gross and Wilczek explicitly worked out the Q^2 -dependence of many of the parton distribution functions and structure functions. An important conceptual as well as technical development in the parton models was the idea of the evolution of parton distribution functions as pioneered by Altarelli and Parisi [80]. Both Gross and Wilczek, and Georgi and Politzer showed how some crucial ingredients of the Altarelli-Parisi equations can be obtained from first principles from QCD based on asymptotic freedom.

A question that may arise naturally at this point is whether all these phenomenological successes are due to QCD or due to field theories possessing AF. An important result towards answering that is due to Coleman and Gross [81] who stated that *no theory which does not involve non-abelian gauge mesons can be asymptotically free*. That certainly seems to be so for $D = 4$. However, $\lambda\phi^3$ theory in $D = 6$ is asymptotically free [82]. Unfortunately $\lambda\phi^3$ theories are sick non-perturbatively! Asymptotic freedom in QCD has allowed for a systematic perturbative treatments to higher order, for example, the pioneering work of Bardeen et al. [83]. Perturbative QCD has grown into a very active and exciting field on its own [84]. But quarks are not liberated even at the highest energies and the true meaning of perturbative QCD is not an obvious one. We shall come to this problem of QCD, the *Quark Confinement Problem* shortly.

19.7 The Static Quark-Antiquark Potential in QCD

In this section we turn to how asymptotic freedom enables one to calculate the static potential between infinitely massive quarks and antiquarks at *short distances*. At larger distances this will be determined by Lattice Gauge Theory techniques. The static potential will be the object of central interest in establishing string-like behaviour in QCD, and its description in terms of Effective String Theories. The role of asymptotic freedom is in enabling perturbative techniques to be applicable. To that extent the situation is exactly as in QED. As already emphasized, the static potential is essentially the two-point function, equivalently, the propagation function. The pioneering works of Gell-Mann and Low [45] have shown how the powerful methods of the renormalization group(RG) work towards this goal.

So similar ideas should work in determining the short-distance behaviour of the static potential in QCD. At short distances, equivalently at high momentum transfers, asymptotic freedom guarantees the smallness of the coupling necessary for a perturbative analysis. However, important differences crop up in comparison to QED. In QED, the coupling, apart from being small, could also be chosen to be a constant. In QCD, as we have discussed at length so far, couplings can run and slide too. This makes the problem technically more complex. The situation is further aggravated by issues of scheme dependence, gauge-independence, etc.

In addition to schemes like the minimal subtraction, modified minimal subtraction, etc., the static potential gives rise to additional schemes, although, as before, one can relate couplings in one scheme to those in others by finite renormalizations. Coming to the static potential, there are essentially four additional schemes called $V, \bar{V}, F$ and SF [85–87]. Of these we shall make a brief mention of the V and $\bar{V}$ schemes, but discuss the F scheme in some detail as our lattice investigations were based on that [88]. These different schemes arise mainly from the momentum space and coordinate space descriptions on the one hand, and from the use of potentials vis a vis forces on the other. These schemes are defined by ($C_F = \frac{4}{3}$ for QCD):

$$V(r) = -C_F \frac{\alpha_{\bar{V}}(r)}{r} \quad \bar{V}(Q) = -4\pi C_F \frac{\alpha_V(Q)}{Q^2} \quad F(r) = C_F \frac{\alpha_{qq}(r)}{r^2} \quad (19.59)$$

The momentum transfer Q and position r are related via Fourier transforms. Denoting the schemes generically by S , it is clear that the effective couplings α_S can all be expanded in terms of $\alpha_{\overline{MS}}$ discussed previously as:

$$\alpha_S = \alpha_{\overline{MS}} + c_0^S (\alpha_{\overline{MS}})^2 + \dots \quad (19.60)$$

The desirability of one scheme over the other lies in the behaviour of the coefficients c_i^S . In the (Force) scheme, they are somewhat smaller [85]. For the F scheme, the c_0 is given [88] by

$$c_0^F = 8\pi \beta_0 \left(\gamma_E - \frac{35}{60} \right) \quad \beta_0 = \frac{11}{(4\pi)^2} \quad (19.61)$$

for pure SU(3) QCD ($n_f = 0$). γ_E is the Euler-Mascheroni constant with numerical value 0.57722. The two-loop result for $\alpha_{\overline{\text{MS}}}(r)$ is given by

$$\alpha_{\overline{\text{MS}}}(r) = \frac{1}{4\pi\beta_0 f_1} \left\{ 1 - \frac{\beta_1 f_2}{\beta_0^2 f_1} \right\} \quad (19.62)$$

Here β_1 is related to the b_1 introduced earlier, and takes the value $\frac{102}{(4\pi)^4}$ for pure QCD. The functions f_1, f_2 are the position space analogs of the logarithms encountered in the running couplings before; specifically, $f_1 = -\ln(r^2 \Lambda_{\overline{\text{MS}}}^2)$ and $f_2 = \ln f_1$. Thus we have all the ingredients to determine $\alpha_{qq}(r)$ and hence the force $F(r)$.

It will turn out later that rather than either $V(r)$ or $F(r)$, a scaled second derivative $c(r) \equiv \frac{r^3}{2} V''(r) = -\frac{r^3}{2} F'(r)$ is more convenient. It is straightforward to put all the ingredients above to arrive at

$$c(r) = -C_F \{ (\alpha_{\overline{\text{MS}}}(r) + c_0 \alpha_{\overline{\text{MS}}}^2(r)) - \frac{r}{2} \partial_r (\alpha_{\overline{\text{MS}}}(r) + c_0 \alpha_{\overline{\text{MS}}}^2(r)^2) \} \quad (19.63)$$

19.8 Colour Confinement in QCD

Despite the impressive progresses made by perturbative QCD, the fact that neither quarks nor gluons are seen as particles was at one point a major source of scepticism towards the correctness of QCD. In the very early days of the quark models (as, for example, when this author wrote a master's thesis on the quark model in 1966), it was believed this had to do with quarks being very massive, and that it was a matter of time before they would be seen with sufficiently energetic collisions. But till date, when particle collision energies have increased by several orders of magnitude, this has not happened. Rather than giving up on quarks, the thinking has veered around to the rather extraordinary point of view that somehow the dynamics of quarks does not result in their being manifest as asymptotic states.

In a sense that puts a question mark on what even perturbative QCD really means as from our extensive discussions of the S-matrix, scattering amplitudes only make sense as transition matrix elements between asymptotic states. It would be fair to say that even to this day a fully satisfactory resolution of this is not at hand.

This topic is highly technical and the last word on it is yet to be written. Therefore it is beyond the scope of this book to give a proper account of all the concepts and developments. Instead, we shall be content with providing some key references and then explain the salient issues. Of course, this whole discussion is to be placed in the context of this book whose goal is an elucidation of the Yang-Mills flux tubes, seen as effectively accounting for quark confinement, and its effective string theory description. These key references are Kenneth Wilson's program for a systematic study of the *Confinement of Quarks* [89] (he introduced and developed Lattice Gauge Theories. We will have lot more to say on it in the next chapter.), that of Nambu on *Strings Monopoles and Gauge Fields* [90], also his paper on *Magnetic and Electric*

Confinement of Quarks [91], Stanley Mandelstam's Physics Reports article [92], the paper by Nielsen and Olesen on string-like excitations in QFT [93], that by t'Hooft [94], by Parisi [95], Creutz [96], by Jevicki and Senjanovic [97], among many more. The area is so active that there are many conferences solely devoted to it! Jeff Greensite has provided a nice overview in his Lecture Notes on the subject [98]. For the *Dual Superconducting Mechanism*, the review by Ripka is very useful [99].

The first conceptual issue was whether there could be *any* field theoretic basis for certain interacting fields not manifesting in asymptotic particle states. In our discussion on the Kallen-Lehmann representation in Chap. 7, we had pointed out the logical possibility when the interaction strengths are *Maximal*. Wilson in the above-cited [89] highlights the work on D=2 QED by Schwinger [100] wherein the electrons are absent as asymptotic states. Furthermore, the electrons do manifest themselves in deep inelastic processes (see Wilson's paper for further discussions and important references). So the Schwinger model does indeed look like a proper direction to follow for the quark confinement problem, although in all these years there has not been much progress in realizing a four-dimensional analog of the Schwinger model. Another drawback seems to be that in the Schwinger model, gauge bosons do appear as asymptotic states, albeit with mass. But in QCD, not just quarks, even gluons are confined.

The thinking further veered towards the view that the confinement of quarks is *absolute* in the sense that no amount of energy could liberate the quarks from hadrons. This hypothesis is certainly incredulous and the focus shifted to answering whether *any* QFT was capable of achieving it. In our earlier discussion of asymptotic freedom, it was pointed out how many thought the infrared slavery at very low energies would be a natural corollary. But clearly such thinking was heuristic at best with no clear demonstration of anything like that happening. The reader is referred to Ref. 1 of [89] for some early proposals to solve the confinement problem.

That in principle field theories could exhibit absolute or permanent confinement came from the tantalizing suggestions of the *Dual Superconductor Mechanism*. The idea was simple and very appealing. To appreciate this consider a superconducting medium (we shall not go into nuances such as type I versus type II superconductors). Due to the Meissner-Ochsenfeld effect, magnetic fields tend to be expelled from the superconducting region. So any attempt to place a *magnetic monopole* will cost a huge amount of energy (infinite, if the superconducting medium extends everywhere) as magnetic flux conservation cannot make the flux vanish everywhere and the superconductor can not tolerate even tiny amount of flux! So, if superconductivity pervades everywhere, magnetic monopoles cannot exist in free state, equivalently, asymptotic states. Now imagine placing a monopole-antimonopole pair. Again, the superconductor will try to minimize regions of magnetic flux. This can be achieved by all the flux being squeezed into a very thin tube. Flux conservation would then imply that the energy per unit length of the flux tube is constant. Consequently it will take infinite amount of energy to completely separate the monopole-antimonopole pair. Equivalently, magnetic charge gets permanently confined.

Coming to quarks and QCD, it is the *colour electric charge* that should be confined. This requires an interchange of electric and magnetic aspects of the superconductor example above. Such an interchange of electric and magnetic fields is called a *Duality Transformation* (care should be taken not to confuse Duality of the dual resonance models). Therefore, if the above picture is correct, the mechanism for quark confinement would be a *Dual Superconductor* [92]. So the big question now becomes whether QCD as a theory displays the dual superconductor behaviour. Since the superconducting “medium” has to extend everywhere, it is clear that it is the QCD vacuum that has to realize this.

In BCS-type superconductors, it is the condensation of pairs of electrons, the *Cooper Pairs*, that is responsible for the superconducting behaviour. By the Duality map, the dual superconductor mechanism must involve condensation of magnetic monopoles. Magnetic monopoles are not explicit in QCD. But one can come up with gauge-field configurations that are monopole like. Whether such configurations are dominant enough to produce dual superconductivity is a very difficult technical question that only *non-perturbative* approaches to QCD can address. To date, impressive progress has been made towards answering such questions based on Lattice Gauge Theory techniques. For various important technical details the reader is referred to [99].

The picture that is directly of interest to us in this book is the thin string-like flux tube connecting the quark-antiquark pair. It reveals itself via the static quark-antiquark potential we discussed before. Therefore another issue that comes up concerns string-like excitations in field theories. A pioneering work in this regard was that of Nielsen and Olesen [93]. Nambu, in his paper [90] makes several interesting observations about the Nielsen-Olesen work which bring out several of the aspects mentioned. Nielsen and Olesen considered the so-called *Abelian Higgs Model*, and explicitly constructed vortex-like solutions. As remarked by Nambu one could view their results in two apparently distinct ways: one as a Higgs effect in an Abelian theory and the other as the Landau-Ginzburg theory of superconductivity with the vortex excitation identified with an Abrikosov flux line (what we just discussed as a means of confining quarks, but in the Dual picture). Nambu further investigates the consequences of considering the Nielsen-Olesen strings which are of finite length and open-ended (Nielsen and Olesen only considered infinite strings or loops). Rather remarkably, Nambu argues that the quarks of the dual (the other type of duality!) models, which are bound by the dual strings, will act as sources of *magnetic* charge! Several seeds of the dual superconductor model are already there in this paper of Nambu.

The other aspect of the Nielsen-Olesen work highlighted by Nambu is the parallel between Abrikosov flux lines and dual strings. We shall show in the later chapters of this book the amazing confluence of all these ideas, namely, quark confinement, flux tubes and bosonic string theory, all emerging out of the non-perturbative investigations of QCD!

19.9 QCD and Chiral Symmetry

In Chap. 18 we have treated this very important aspect of strong interactions in detail. We will therefore be very brief here. The purpose of this subsection is mainly to make this chapter on QCD self-contained with regard to its salient aspects. Even in this chapter we have already mentioned the important role by this concept in the very creation of QCD.

Before the advent of QCD, there were some striking puzzles regarding strong interactions. Among the earliest to draw attention was the surprisingly light mass of the pions which is about a seventh of the nucleon mass. An early picture of the pion was as a bound state of nucleon and antinucleons. While an extremely strong binding could in principle result in a very light pion, this was nevertheless very puzzling. The other puzzling aspect was with the axial vector current and its possible conservation. In particular, the issue was whether the axial vector current, which in many ways appeared to be on par with the conserved vector current could also be conserved. Another puzzle had to do with the remarkable phenomenological success of the Goldberger-Treiman relation $f_\pi g_{\pi NN} = 2 m_N g_A$ [101,102]. Combined with all this was also a desire to understand the almost exact isospin invariance.

In a path-breaking paper [103] Nambu in one go resolved all the difficulties. His major observation was that even the axial current could be exactly conserved provided there exists a *massless* isospin one particle that couples to both the nucleons as well as the axial current. His bold suggestion was that the pion of the real world, which while very light is not really massless, should be identified with this massless particle when certain symmetries are broken in the real world. Consequently, he argued that in the real world, the axial current is only partially conserved. This hypothesis of PCAC or *Partially Conserved Axial Current* subsequently proved to be enormously successful.

Immediately after Nambu's paper, Goldstone [104] stated that when continuous symmetries are broken spontaneously (in modern parlance), there will always be such massless particles. A year later, in 1962, this was proved within QFT by Goldstone, Salam and Weinberg [105]. The symmetry in question came to be known as *Chiral Symmetry*. In such situations, while the equations of motion are symmetric, the ground state is no longer so.

With these remarks, let us now turn to their possible implications for QCD. The most important is that the vacuum state of QCD must spontaneously break chiral symmetry. This typically means that QCD dynamically generates vacuum expectation values that imply such breaking. This can happen even in the ideal limit of massless pions, with the consequent *exact* conservation of the axial vector current. For QCD, that would imply the situation when quarks are massless. The non-vanishing quark masses imply an *explicit* breaking of chiral symmetry too, giving rise to non-zero masses for pions. That the real-world pions are light would then be a reflection of finite, but small, masses of the u , d quarks. Specifically, the QCD vacuum is expected to yield $\langle \bar{q} q \rangle \neq 0$ at least for both u and d quarks. Much before QCD, already in 1968, Gell-Mann, Oakes and Renner [106] had proposed the relation

$$m_\pi^2 f_\pi^2 = -\langle \bar{u} u \rangle (m_u + m_d) \quad (19.64)$$

Since spontaneous breakdown of chiral symmetry can in principle occur even when quarks are massless, it is desirable to seek its consequences in the massless quark case. Indeed there are such (see Leutwyler's lectures on Chiral Perturbation Theory and Lattice QCD [107]). With the defining relations

$$\langle 0 | \bar{u} \gamma^\mu \gamma_5 d | \pi^- \rangle = i\sqrt{2} f_\pi p^\mu \quad \langle 0 | \bar{u} i \gamma_5 d | \pi^- \rangle = \sqrt{2} G_\pi \quad (19.65)$$

it follows that:

$$F_\pi G_\pi = -\langle 0 | \bar{u} u | 0 \rangle \quad (19.66)$$

The smallness of the u , d masses also gives a natural explanation for the observed almost exact isospin invariance. In fact, they do so for the entire $SU(2) \times SU(2)$ invariance of the low-energy hadronic world.

It is to be emphasized that proving all these properties analytically within QCD is still a formidable task. But increasingly accurate and reliable results from Lattice Gauge Theories have shown their essential correctness. The same can be said of Quark Confinement too.

We end this chapter on Quantum Chromodynamics with a very apt quote from Weinberg's book [43] (end of Sect. 19.4) which sums it all nicely: *One of the reasons for the rapid acceptance of quantum chromodynamics in 1973 as the correct theory of strong interactions was that it explained $SU(2) \times SU(2)$ symmetry as a simple consequence of the smallness of the u, d quark masses.*

References

1. H. Fritzsch, *The Development of Quantum Chromodynamics*, Preprint-MPI-PAE-PTh3/84, 1984
2. H. Fritzsch, *The History of QCD*, CERN Courier, 27 Sep 2012
3. Norman Dombey, *Gell-Mann in his own words*, 2020
4. S. Gasiorowicz, *Elementary Particle Physics*, John Wiley and Sons, 1966
5. A. Pais, Phys. Rev. **86**, 663 (1952)
6. M. Gell-Mann, Phys. Rev. **92**, 833 (1953)
7. T. Nakano, K. Nishijima, Prog. Theor. Phys. **10**, 581 (1953)
8. M. Gell-Mann, A. Pais, Phys. Rev. **97**, 1387 (1955)
9. S. Sakata, Prog. Theor. Phys. **16**, 686 (1956)
10. M. Gell-Mann, in *Eightfold Way*, W.A. Benjamin, 1964
11. Y. Neeman, Nuc. Phys. **26**, 222 (1961)
12. V.E. Barnes et al Phys. Rev. Lett. **12** p. 204 (1964); G.S. Abrams et al Phys. Rev. Lett **13** p. 670 (1964)
13. N.D. Hari Dass, *In search of QCD strings*, Indian. J. Phys. **95**(8) p. 1591 (2021)
14. M. Gell-Mann, Phy. Lett. **8**, 214 (1964)
15. G. Zweig, CERN TH 401 & 410 (1964)
16. G. Zweig, *Origins of Quark Model* CALT-68-805 (1980)
17. Kerson Huang, *Quarks, Leptons and Gauge Fields*, World Scientific Publishing, (1982)
18. O.W. Greenberg, Phys. Rev. Lett. **13**, 598 (1964)
19. M. Han and Y. Nambu, Phys. Rev. **B139** p. 1006 (1965)

20. R. Miskimen, *Neutral Pion Decay*, Ann. Rev. Nucl. Part. Sci. 2011/61
21. B. de Wit and J. Smith, *Field Theory in Particle Physics*, North-Holland Personal Library, 1986
22. J. Schwinger, Phys. Rev. **82**, 664 (1951)
23. S.L. Adler, Phys. Rev. **177**, 2426 (1969)
24. W.A. Bardeen, H. Fritzsch and M. Gell-Mann, in *Scale and Conformal Symmetry in Hadron Physics*, John Wiley and Sons, p. 139 (1973); H. Fritzsch, Fortschritte der Physik 22 p. 407 (1974)
25. J. Schwinger, Ann. Phys. **2**, 407 (1957)
26. J.J. Sakurai, Ann. Phys. **11**, 1 (1960)
27. Y. Nambu, A Systematics of Hadrons in Subnuclear Physics, in *Preludes of Theoretical Physics*, ed. by A. de Shalit (North Holland, 1966)
28. J.C. Pati, A. Salam, Phys. Rev. D **8**, 1240 (1973)
29. H. Fritzsch and M. Gell-Mann, in Proceedings of the XVI Int. Conference on High Energy Physics, Chicago, 1972
30. H. Fritzsch, Fortschr. Phys. **22**, 407 (1974)
31. H. Fritzsch, M. Gell-Mann, H. Leutwyler, Phys. Lett. **47B**, 365 (1973)
32. S. Weinberg, Phys. Rev. Lett. **31**, 494 (1973)
33. D.J. Gross, F. Wilczek, Phys. Rev. D **8**, 3633 (1973)
34. J.C. Pati, A. Salam, Phys. Rev. Lett. **36**, 11 (1976)
35. G. Rajasekaran and P. Roy, Pramana 5 p. 303 (1975); Phys. Rev. Lett 36 p. 355 (1976)
36. Saurabh D. Rindani, *How do we know the charges of quarks*, [ArXiv:0210054v1](https://arxiv.org/abs/0210054v1) [hep-ph](2002)
37. H. Kendall, in *Proceedings of the 1971 International Symposium on Electron and Photon Interactions*, Ed. N.B. Mistry, 1972
38. J.D. Bjorken, Phys. Rev. 148 p. 1467 (1966); Phys. Rev. 179 p. 1547 (1969)
39. R.P. Feynman, Phys. Rev. Lett. **23**, 1415 (1969)
40. R.P. Feynman, *Photon-Hadron Interactions* (Benjamin, New York, 1972)
41. C.N. Yang, R. Mills, Phys. Rev. **96**, 191 (1954)
42. Oskar Klein, in *New Theories of Physics*, Warsaw Conference, 30 May-3 June 1938
43. Steven Weinberg, *Quantum Theory of Fields-II*, Cambridge University Press
44. E.C.G. Stueckelberg, A. Peterman, Helv. Phys. Acta **26**, 499 (1953)
45. M. Gell-Mann, F.E. Low, Phys. Rev. **95**, 1300 (1954)
46. Claude Itzykson and Jean-Bernard Zuber, *Quantum Field Theory*, McGraw Hill Publishers
47. J.D. Bjorken and Sidney Drell, *Relativistic Quantum Fields*, McGraw Hill Publishers
48. N.N. Bogoliubov, D.V. Shirkov, *Introduction to the theory of Quantized Fields* (Interscience, New York, 1959)
49. C.G. Callan, Phys. Rev. D **2**, 1541 (1970)
50. C.G. Callan, S. Coleman, R. Jackiw, Ann. Phy. (NY) **59**, 42 (1970)
51. K. Symanzik, Comm. Math. Phys. **18**, 224 (1970)
52. K.G. Wilson, Phys. Rev. B4 p. 3174,3184 (1971)
53. K.G. Wilson, Rev. Mod. Phys. **47**, 773 (1975)
54. M. Sivakumar, *Introductory Quantum Field Theory*, Ane Books Pvt. Ltd
55. G.t'Hooft and M. Veltman, Nuc. Phy. B44 p. 189 (1972)
56. W.E. Caswell, F. Wilczek, Phys. Lett. B **49**, 291 (1974)
57. G.t'Hooft, Nuc. Phy. B61 p. 455 (1973)
58. S.D. Joglekar, Phys. Rev. D35 p. 759 (1987); Pramana 34 p. 91 (1990)
59. G.t'Hooft, unpublished, announced at the International Conference on Yang-Mills Fields, Marseilles, 1972
60. D.J. Gross, F. Wilczek, Phys. Rev. Lett. **30**, 1343 (1973)
61. H.D. Politzer, Phys. Rev. Lett. **30**, 1346 (1973)
62. D.J. Gross, F. Wilczek, Phys. Rev. D **8**, 3633 (1973)
63. D.J. Gross, F. Wilczek, Phys. Rev. D **9**, 980 (1974)
64. M. Bace, Phys. Lett. B **78**, 132 (1978)
65. W.E. Caswell, Phys. Rev. Lett. 33 p. 244 (1974)

66. D.R.T. Jones, *Nuc. Phys.* **B75**, 531 (1974)
67. G. Egorian, O.V. Tarasov, *Theor. Math. Phys.* **41**, 883 (1974)
68. O.V. Tarasov, V.A. Vladimirov, A.Y. Zharkov, *Phys. Lett. B* **93**, 429 (1980)
69. J.A.M. Vermaseren, arXiv-math-ph 0010025
70. S.A. Larin, J.A.M. Vermaseren, *Phys. Lett. B* **303**, 334 (1993)
71. T. Ritbergen, J.A.M. Vermaseren, S.A. Larin, *Phys. Lett. B* **400**, 379 (1997)
72. M. Czakon, arXiv 0411261v2(hep-ph)
73. N. Brambilla, A. Pinoda, J. Soto, A. Vairo, *Phys. Rev. D* **60**, 091502 (1999)
74. Y. Schröder, *The static potential in QCD*, PhD Thesis, DESY-THESIS-1999-021, June 1999; *Phys. Lett. B* 447 p. 321 (1999)
75. E.D. Bloom et al *Phys. Rev. Lett.* 23 p. 930 (1969)
76. K. Wilson, *Phys. Rev.* **179**, 1499 (1969)
77. K. Wilson, *Phys. Rev. D* **3**, 1818 (1971)
78. D. Wilson, *Proceedings of the Kiev Conference* (1970)
79. H. Georgi, H.D. Polizer, *Phys. Rev. D* **9**, 416 (1974)
80. G. Altarelli, G. Parisi, *Nuc. Phys.* **B126**, 298 (1972)
81. S. Coleman, D.J. Gross, *Phys. Rev. Lett.* **31**, 851 (1973)
82. A.J. Macfarlane, G. Woo, *Nuc. Phys.* **B77**, 91 (1974)
83. W.A. Bardeen, A.J. Buras, D.W. Duke, T. Muta, *Phys. Rev. D* **18**, 3998 (1978)
84. T. Muta, *Foundations of Quantum Chromodynamics: An Introduction to Perturbative Methods in Gauge Theories*, Lecture Notes in Physics Vol 78 (3rd Ed.), World Scientific Publishers
85. S. Necco, R. Sommer, *Phys. Lett. B* **523**, 135 (2001)
86. S. Melles, *Phys. Rev. D* 074019 (2000)
87. A. Laschka, N. Kaiser and W. Weise, IX International Conference on Quark Confinement and Hadron Spectroscopy(QCHS), Madrid, 2010
88. N.D. Hari Dass and Pushan Majumdar, *JHEP* 0610 p.020 (20006)
89. K. Wilson, *Phys. Rev. D* **10**, 2445 (1974)
90. Y. Nambu, *Phys. Rev. D* **10**, 4262 (1974)
91. Y. Nambu, *Phys. Rep.* **C23**, 237 (1976)
92. S. Mandelstam, *Phys. Rep.* **C23**, 245 (1976)
93. H.B. Nielsen, P. Olesen, *Nuc. Phys.* **B61**, 45 (1973)
94. G.t'Hooft, *High Energy Physics Editorice Compositori Bologna*, 1975
95. G. Parisi, *Phys. Rev. D* **10**, 870 (1975)
96. M. Creutz, *Phys. Rev. D* **10**, 2696 (1974)
97. A. Jevicki, P. Senjanovic, *Phys. Rev. D* **11**, 860 (1975)
98. J. Greensite, *An Introduction to The Confinement Problem*, Lecture Notes on Physics 821, Springer
99. G.Ripka, arXiv:0310102(hep-ph)
100. J. Schwinger, *Phys. Rev.* 125 p. 397 (1962); *Phys. Rev.* 128 p. 2425 (1962)
101. M.L. Goldberger, S. Treiman, *Phys. Rev.* **110**, 1178 (1958)
102. M.L. Goldberger, S. Treiman, *Phys. Rev.* **111**, 354 (1958)
103. Y. Nambu, *Phys. Rev. Lett.* **4**, 380 (1960)
104. J. Goldstone, *Nuovo Cimento* **9**, 154 (1961)
105. J. Goldstone, A. Salam, S. Weinberg, *Phys. Rev.* **127**, 965 (1962)
106. M. Gell-Mann, R.J. Oakes, B. Renner, *Phys. Rev.* **175**, 2195 (1968)
107. H. Leutwyler, *Chiral Perturbation Theory and Lattice QCD* (School on Quantum Field Theories on Lattice, Saha Institute of Nuclear Physics, Kolkata, 2013)

20.1 Introduction

Many crucial aspects of QCD like *Colour Confinement*, *Spontaneous Breaking of Chiral Symmetry*, and even the observed spectrum and interactions of strongly interacting particles are beyond the reach of perturbative techniques which were so successful in the spectacular developments of QED and electroweak unified theories. Among many issues, one that assumes centre stage is a *non-perturbative regularization* of field theories in general, and of Quantum Chromodynamics (QCD) in particular. In the Lattice regularization, the continuum nature of space-time in the Euclidean formulation, and of space in the Hamiltonian formulation, is replaced by a discrete set of points. The inverse of the lattice spacing, in some average sense, plays the role of an ultraviolet cut-off. Infrared divergences that can also plague continuum field theories are not circumvented this way, and have to be dealt with separately. It is worth emphasizing at this point the fundamental difference between lattice regularized continuum field theories, and the naturally occurring lattice theories of the condensed matter kind; in the latter, the lattice spacings have a physical meaning.

In what follows, we shall mostly be interested in the lattice formulations of *Gauge Theories*. Before proceeding further we briefly recapitulate the history of Lattice Gauge Theories (LGT's). The first discussion of theories closest to them is to be found in Franz Wegener's work on Duality in what he called "Generalized Ising Models" [1], though at that time its relation to LGT's, as they are understood today, was not obvious. Nevertheless, Wegner should be credited with the invention of Lattice Gauge Theories.

We shall summarize the genesis and broad aspects of Wegner's pioneering work. The reader is encouraged to read the original paper which is both lucid and elaborate. What seems to have motivated Wegner is the desire to find systems where phase transitions occur without associated *local order parameters*. He arrives at it by enlarging the concept of *Duality* in Ising systems first introduced by Kramers

and Wannier [2]. While the original Kramers-Wannier duality was between the high and low temperature behaviours of the same system, the 2-d Ising model, Wegner introduced more general dualities connecting *different* Ising systems. Specifically, he found a duality between the $d = 3$ Ising model and a $d = 3$ model with Ising variables $\sigma_{\mathbf{i},\mu}$, living on the *links* which are, for example, edges connecting a lattice site $\mathbf{i}$ to its nearest neighbours in the direction μ (unlike the $d = 3$ Ising model where such variables live on the *sites*). The Hamiltonian of the model is invariant under *gauge transformations* of the type

$$\sigma_{\mathbf{i},\mu} \rightarrow g_{\mathbf{i}} \sigma_{\mathbf{i},\mu} g_{\mathbf{i}+\hat{n}_{\mu}} \quad (20.1)$$

where the g 's are also Ising variables, but now living on sites. The connection of this to the gauge transformations of QED and QCD, discussed, for example, in Chap. 19, is not obvious at this stage. That such transformations are actually gauge transformations on lattices will become clear after we have developed the rest of this chapter. Stated more precisely, Wegner's new class of Ising models are LGT's where both the link variables and the gauge transformations are elements of the discrete, abelian, gauge group Z_2 . In contrast, the gauge group of QCD is the continuous (Lie) non-abelian group $SU(3)$, and that of QED is the continuous but abelian group $U(1)$. The corresponding local gauge transformations have been discussed extensively in Chap. 19.

The other notable feature of Wegner's LGT is that only space is discretized, in contrast to the Euclidean LGT's to be discussed shortly. This had to do with the fact that Wegner's focus was on equilibrium statistical mechanics where time has no role to play. Consequently, Wegner developed a *Hamiltonian Formalism* for Z_2 -LGT. The Hamiltonian formalism for LGT's with continuous groups, relevant for QCD, for example, was developed by Kogut and Susskind [3,4] much later, a year after Kenneth Wilson's pioneering work on Lattice Gauge Theories in 1974 [5], based on the Euclidean approach. In the Hamiltonian approach, the field variables are still *operator valued*. For the purposes of this book, the focus will be entirely on Euclidean LGT's.

Rather remarkably, Wegner's construction explicitly discussed all the essential features of LGT's like Plaquette action, the Wilson loops, as well as their dual counterparts, the t'Hooft loops. It even anticipated the important criteria for confinement like the area law versus the perimeter law, etc. As remarked by Erhard Seiler in his wonderful lecture notes [6], even Wilson, the pioneer of LGT's was apparently unaware of Wegner's great contributions!

The gauge group addressed by Wegner was the simplest kind of group one could envisage, namely, the abelian discrete group Z_2 , while the gauge group relevant for QCD is the non-abelian group $SU(3)$. Kenneth Wilson's pioneering paper on LGT [5] was strongly motivated by the desire to understand the vexing problem of Quark Confinement. Nevertheless, as suggested by the works of Mack and Petkova [7] on the one hand, and that of Frohlich [8], there is an intimate connection between the confining properties of Z_N and $SU(N)$ gauge theories! The picture that was suggested was that $SU(N)$ LGT's confine if Z_N theories do! (see [6] for a discussion).

Three years after Wegner's work, Kenneth G. Wilson independently invented LGT's in their modern avatar, motivated primarily by a way to tackle the quark confinement problem non-perturbatively [5]. In fact the title of the paper was *Confinement of quarks*. In this paper, Wilson gave a systematic construction of Euclidean LGT's for QCD. Here he solved the problem of finding discretized version of the QCD action as well as the discretized form of the non-abelian gauge transformations under which the action was *exactly* invariant. This necessitated introducing the gauge fields as elements of the gauge group $SU(3)$ living on the links of the lattice, exactly as in Wegner's LGT. In mathematical terms, the link variables are elements of *holonomy*. We shall explain these non-trivial aspects of Wilson's LGT shortly. The quark fields, on the other hand, were to live on the sites of the lattice. All continuum space-time derivatives were to be replaced by *finite differences*. This was also the case in Wegner's theory though in that case only spatial derivatives were involved. One of the complications is that Dirac fields, in order to be consistent with the Pauli exclusion principle or, equivalently, the Fermi-Dirac statistics, have to be represented by anti-commuting Grassmann variables as had already been discussed by Berezin in [9], and this feature was already necessary for a path-integral quantization of Dirac fields. A deeper problem arises for the Dirac fields owing to the fact that the space-time derivatives occur *linearly* in their action. This leads to the notorious problem of *species doubling*. In what follows, we shall not concern ourselves with fermionic fields on the lattice. We shall, however, give a list of excellent books for the interested reader to follow this aspect of LGT's. Wilson, however, does not address this problem in his pioneering work. From the list of references in Wilson's paper, it is also clear that he was apparently unaware of Wegner's earlier work.

It is worth emphasizing at this stage that even after Wegner's work, LGT's were independently discovered (invented) by Smit [10] and Polyakov [11] even before Wilson's work. This is not very well known. Jan Smit discovered LGT as a graduate student at UCLA during 1972–73; he, however, did not publish his findings. It is said that he encountered the fermion doubling problem to which he did not find a satisfactory solution at that time. The unpublished documentation was sent as private communication to Wilson by Robert J. Finkelstein, who was Jan Smit's thesis advisor. Polyakov also did not publish his ideas on lattice gauge theories. He however mentions his unpublished work in [11] in which he also applies lattice gauge techniques (compact abelian groups) to address the infrared problems in QFT's. Though the early works of Jan Smit and Polyakov remained unpublished, they have been explicitly acknowledged by Wilson himself [12]; quoting verbatim his historical note there, *the first studies of lattice gauge theories were carried out independently by Wegner, Smit, Polyakov, and finally myself. The original feature of my paper is its discussion of quark confinement in the lattice theory for strong coupling*.

Apart from a non-perturbative regularization of quantum field theories, the lattice approach delivered some outstanding advantages! The most notable among them were the manifest gauge invariance, an exact mapping of QFT's to *classical* statistical mechanical theories, albeit in one higher spatial dimensions (see [6] for a discussion of this fundamental connection), and the ability to simulate, to arbitrary accuracy (at least in principle), on computers. Another notable feature of the lattice regularization

that emerged from the works of Osterwalder and Seiler [13] as well as of Lüscher [14] was that for every value of the lattice spacing (cut-off), a positive-definite space of the physical states of the corresponding gauge-field theory can be constructed.

Over the last four decades or more, these features have contributed significantly to our understanding of quantum field theories in general, and of QCD and the standard model in particular. Nearly four decades of the proceedings of the International Symposia on Lattice Field Theories are an eloquent testimony to this.

There are three essential conceptual pillars to the framework of *Euclidean* lattice field theories, and we now turn to their brief but self-contained elaborations, along with some remarks on their logical and structural independence. The first of these is Richard Feynman's *Path-Integral* approach to Quantum Mechanics, and subsequently, to Quantum Field Theory [15]. The central idea here was to do away with the operator approaches to describe quantum phenomena, thought till then to be the very essence of Quantum Theory, and instead use only the so-called *c-number* functions. This is at the heart of computer simulations of lattice field theories. The price one had to pay for this was that entire *histories*, or more precisely *paths*, had to be considered. Since the other essence of quantum theory is the renunciation of *space-time trajectories*, path-integrals were mystifying on that count too. It took some time to appreciate that the paths in a path-integral had little to do with space-time trajectories dynamically determined as in classical theories. Fermions do pose problems, however. That has to do with the fact that momentum appears quadratically in the Schrödinger equation while it appears linearly in the Dirac equation on the one hand, and with the fact that fermions have to obey Pauli exclusion principle. The latter essentially means there is no way to represent fermion fields by pure numbers.

In fact, many path-integral representations may be possible of which the so-called *configuration space*, and, *phase space* representations are the most explored. The path-integral representation of the Dirac field is necessarily of the latter kind. There are many excellent resources on path-integral quantization. The book by Feynman and Hibbs [16] is a must. It has become textbook material now and almost every book on QFT is based on them. But most of them tend to present path-integral approach, also called *Functional Methods* by some, in an uncritical manner as finished products. We recommend the reader to Weinberg's book [17] (Chap. 9) for a detailed yet critical perspective. He also discusses the path-integral approach from the perspectives of canonical quantization. The Quantum Field Theory book by Itzykson and Zuber also has nice and extensive discussions. I have discussed some fundamental limitations of Feynman path-integral issues in [18]. Path-integral methods have found wide applicability beyond QFT in many diverse areas. For an encyclopaedic coverage the reader is referred to the book by Kleinert [19]. There are even applications in Finances as can be found in Belal Baaquie's book [20].

The path-integral formulation has also had a big impact on the development of string theories through Polyakov's pioneering works [21]. For a detailed discussion of this approach we refer the reader to Polchinski's book on string theory [22]. This will have ramifications for effective string theories also as will be elaborated in the concluding chapters of this book.

The next ingredient is a rather mysterious aspect of quantum field theories called *Euclideanization*. Logically, it is a structure *independent* of the path-integral aspect. Its implications have been extensively discussed even in the continuum formulation of quantum field theories. The Euclideanization procedure and the underlying deeper aspects were first formulated by Schwinger in [23], and shortly afterwards by Tadao Nakano in [24]. There has been explosive activity subsequently; we shall try to give an overview of the essentials.

Before the advent of special relativity, the view of space-time was that of absolute space on its own, and, of absolute time on its own. Mathematically speaking, the manifold was thought to be $R^3 \times T$. This was so both of Newtonian mechanics and of non-relativistic quantum mechanics. The relevant relativity principle was that of *Galilean Relativity*. Space and time appeared asymmetrically, as can be seen directly from Newton's laws or from Schrödinger equation (or equally from Heisenberg's equations). With the advent of special relativity and the Lorentz Group (Poincare Group), space and time were put on a symmetrical and unified footing. Nevertheless, the nature of space and time is indeed very different, as we all know. In more geometric terms, the "metric" representing the space-time of special relativity is the *Minkowskian Metric* (in Cartesian-like coordinates) $ds^2 = -c^2 dt^2 + d\mathbf{x}^2$, with c the velocity of light in vacuum (this is a matter of convention; as has been amply stressed throughout the book, there are many different conventions but with unambiguous prescriptions to relate them. Physics is, of course, unaffected by the different choices. Often, the choice of units is so made that $c = 1$). The fundamental difference between space and time is manifested through the differing signs in the respective metric contributions. This is what makes an invariant decomposition of events into *past*, *present* and *future* meaningful. This is also called the *causal structure* of space-time, and is a pre-requisite for *Causality* which has played such a major role in our deliberations so far.

There is a closely related four-dimensional space (the generalization to other dimensions is straightforward), called *Euclidean space*, described by the metric (again in Cartesian coordinates) $ds_E^2 = \sum_{i=1}^4 dx_i dx_i$. Clearly, the geometry and topology of the Minkowskian and Euclidean spaces are very different, the great similarities between their metrics notwithstanding. The differential equations for "wave" propagation are also fundamentally different, with hyperbolic differential equations in Minkowski case and elliptic differential equations in Euclidean case.

It is against this backdrop that Schwinger's proposal of the *Euclidean Postulate* [23, 25] seemed so extraordinary! The essence of this postulate (see p. 44 of [26]) is that the amplitudes of a relativistic quantum field theory continue to be meaningful and invariant under the mapping of Minkowski space onto the Euclidean space. As stated by Schwinger in [23], *a detailed correspondence can be established between RQFT and a mathematical image based on 4-d Euclidean manifold*. He expressed this in a more picturesque way by saying it was as if nature formulated her thoughts first in the Euclidean language and then rendered them into the Minkowskian! On a more technical note (to those familiar with axiomatic approaches to QFT), the Euclidean Postulate is the detailed correspondence that can be established between the *Wightman Functions* of Minkowskian QFT and the *Schwinger Functions* of the

Euclidean analog. The reader is referred to the excellent book by Haag [27] for a clear discussion of the Wightman functions.

Schwinger also offers a useful group theoretic perspective on the Euclidean Postulate. Even before his work, mathematicians knew that some representations of the Lorentz group could be obtained from the corresponding Euclidean group representation through the so-called Weyl's *unitary trick*. Schwinger goes on to claim that *all* representations of the Lorentz group that are of physical interest can be obtained this way. Essentially due to the causal structure of the Minkowski space(time), Green's functions of the Lorentzian description are intrinsically complex quantities, and there are two linearly independent sets of them. Schwinger then states that it is an indication of the simplification obtained through Euclideanization that completely *real* Green's functions can be defined.

It took nearly 15 years after the works of Schwinger and Nakano before Osterwalder and Schrader wrote their seminal paper *Axiom's for Euclidean Green's Functions* [28, 29]. In this they worked out conditions that Euclidean correlation functions (Schwinger functions mentioned above) must obey in order they are equivalent to the Minkowski correlation functions (Wightman functions). In particular, they introduced the powerful concept of *Reflection Positivity* which the Euclidean correlation functions must satisfy. In essence, reflection positivity in Euclidean QFT's guarantees a positive-norm Hilbert space and unitarity of the corresponding Minkowskian QFT's. There are difficulties with fermionic fields even with regard to Euclideanization, as has been clearly brought out by Osterwalder and Schrader in [28, 29]. One set of difficulties is easy to identify and has to do with the fact that properties of γ -matrices like hermiticity are intimately tied up with the metric (their defining relations are $\gamma_\mu \gamma_\nu + \gamma_\nu \gamma_\mu = 2 g_{\mu\nu}$). More technically, these have to do with the representations of *Clifford Algebras* and their intimate dependence on both dimensions and metric. For an extensive discussion the reader is recommended to read Appendix E of [30]. Other difficulties with Euclideanized fermion fields are not so straightforward to grasp; they involve the *Grassmannian* nature of these fields.

One of the great virtues of Euclideanization is actually for the path-integral formulation itself, and this will turn out to be absolutely crucial for Lattice Field theories. To appreciate this important aspect, let us take a schematic look at the Feynman path-integral for a QFT with fields $\phi(x)$ (generalizations to more complicated QFT's are not hard):

$$\mathcal{Z} \equiv \int \mathcal{D}\phi(x) e^{i S_{cl}(\phi)} \quad (20.2)$$

There are several reasons because of which this can at best be a very formal expression, mathematically speaking. An important one is that this is an *infinite-dimensional* integral with a highly *oscillatory* integrand! Euclideanization of QFT's on *Flat* space-time manifold at least fixes this big lacuna. In the process of Euclideanization, $x^0 = ct \rightarrow -i x_4$ (the sign here is extremely important and must be treated with care), and the Minkowskian fields are replaced by Euclidean fields. The space-time metric goes from $ds_M^2 = -c^2 dt^2 + dx_i \cdot dx_i$ to $ds_E^2 = \sum_{i=1,4} (dx_i)^2$. The four-dimensional volume element $d^3x dt$ goes to $-i \prod_{i=1}^4 dx_i$.

To see what happens to $S_{cl}(\phi)$ under Euclideanization, let us illustrate with the example of a scalar field. The reader is urged to consult more advanced resources (a list of which will be given shortly) to see how this works for generic QFT's in flat space-time. There are some issues with fermionic fields to which we shall return soon. For the metric in consideration, the Lagrangean density can be taken to be (again, more complicated cases with derivative couplings can also be treated)

$$\mathcal{L} = \frac{1}{2} \{ (\partial_t \phi)^2 - \nabla \phi \cdot \nabla \phi \} - V(\phi) \quad (20.3)$$

Again, the form of this depends on the metric convention, which we have chosen to be $\eta^{\mu\nu} = \text{diag}(-1, 1, 1, 1)$. The sign of the potential $V(\phi)$ is dictated by the physical consideration that the Hamiltonian must be bounded from below for vacuum stability. In the more complicated cases (like those involving derivative couplings, theories with fermions, etc.) too, it is this consideration that is crucial. With the present metric conventions, $V(\phi) \geq 0$. Upon Euclideanization, one gets

$$\mathcal{L}_E = - \left\{ \sum_{i=1}^4 (\partial_i \phi)^2 + V(\phi) \right\} \leq 0 \quad (20.4)$$

Consequently, the phase of the integrand of the path-integral $i S_{cl} \rightarrow -S_E$, with the Euclidean action given by

$$S_E = \int d^4x \left\{ \sum_{i=1}^4 (\partial_i \phi)^2 + V(\phi) \right\} \geq 0 \quad (20.5)$$

Hence the troublesome oscillatory integrand $e^{i S_{cl}}$ has been rendered into the highly damped e^{-S_E} !

Hawking sought to extend the methods of path-integrals, and of Euclideanization, to the problems of Quantum Gravity where the space-times involved are generically those with non-trivial geometry and topology. We recommend the reader to Hawking's article *The path-integral approach to quantum gravity* in [31]. However, the level of rigour of Schwinger's as well as that of Osterwalder and Schrader about the equivalences is yet to be reached for general space-times. One difficulty with gravitational fields can be understood straight away: the Einstein-Hilbert action

$$S_{EH} = \frac{1}{16\pi G} \int d^4x \sqrt{-g} R \quad (20.6)$$

even upon Euclideanization depends on the Euclidean scalar curvature R_E which need not be of a definite sign.

We finally come to the last of the three pillars, namely, discretization of the continuum space-time, also called *Latticization*. This amounts to replacing the space-time continuum by a discrete (possibly infinitely many) set of points. The average separation between them, in some suitable sense, acts as the inverse of an ultraviolet cut-off

in momenta, and serves to regularize the field theory. There is a lot of freedom in the choice of discretization. For example, the discrete set could be a regular hypercubic lattice (a frequent and very convenient choice), a triangular lattice in 2-d or its higher simplicial generalization in higher dimension, etc. Even a *random lattice* has been made use of. Even after fixing the lattice, there are many ways of approximating the continuum derivatives of fields by finite differences on the lattice. Put together, one can in principle choose a large number of lattice theories all representing the same (naive) continuum field theory. The end result of putting together the path-integral representation, Euclideanization and discretization is that the partition function $\mathcal{Z}$ of QFT's (more precisely, the vacuum-to-vacuum transition amplitude) looks mathematically *indistinguishable* from the thermal partition function of *classical* statistical mechanics, albeit now in four spatial dimensions. Modulo various nuances of each of the three main steps, this is an *exact* mapping between QFT's in $d + 1$ space-time dimensions and classical statistical mechanical theories in $d + 1$ spatial dimensions.

It is this exact mapping that gives the lattice approach its enormous power. On the one hand, on the analytical side, it allows many powerful techniques of statistical mechanics like *Correlation Inequalities* to be applied to QFT's; an excellent source for a discussion of these is Erhard Seiler's lecture notes [6] which also gives a lucid introduction to the Osterwalder-Schrader positivity as well as many aspects of both continuum QFT's and FT's on lattice. On the other hand, it allows for detailed numerical investigations, again based on the rich repertoire of numerical and computational bag of tricks from statistical mechanics like Monte Carlo simulations, so-called molecular dynamics simulations, etc. It was the pioneering work of Creutz in this direction that opened the floodgates to the numerical simulations of QCD, other field theories and spin systems [32]. Today we owe much of our progress on the non-perturbative aspects of QFT's including QCD to the lattice approach.

With such discretization, several important features may be lost. Most important for QFT's is that Lorentz invariance no longer holds. Instead, depending on the lattice, only invariance under subgroups of the Lorentz group like the hypercubic group may be left. In the case of a *Random Lattice*, even that limited invariance may be gone. The hope is that in the *continuum limit* of the lattice field theories, these important symmetries are restored. But exactly how that may happen is a very thorny technical issue. The topology of space-time is also compromised. Even the topological aspects of the fields are also affected. Nevertheless, with ingenuity, many of these features can be recovered. The description of fermionic fields is also drastically affected with conceptual difficulties like *species doubling* posing serious challenges. This is deeply rooted, on the one hand, in the topology of the momentum space (the so-called Brillouin Zone) being changed to that of torii. A related conceptual difficulty, on the other hand, is that of putting *chiral fermions* on the lattice. Considering their importance in the construction of the *Standard Model* of particle physics, the fermions on the lattice pose significant challenges. An overview can be found in my lectures *Regularization of Chiral Gauge Theories* [33].

In continuum QFT's, first one regularizes the theory by introducing a cut-off (though at the level of regulating the relevant Feynman integrals). Then one carries out the *renormalization* procedure, after which there is still some residual cut-off

dependence, which is removed in the last step. In the lattice, this last step is achieved through the so-called *statistical scaling*, which is highly non-trivial. We shall explain this in the remainder of this chapter.

In 1978, 4 years after Wilson's work, Osterwalder and Seiler, in a very influential work, extended the Osterwalder-Schrader work to Euclidean LGT's [13]. In particular, they formulated the Osterwalder-Schrader criteria to LGT's, essential for a Hilbert space of physical states with a positive-definite metric. In addition, the paper contains many important discussions like on confinement and the Higgs mechanism. The treatment includes fermionic fields also. Martin Lüscher gave an independent demonstration of how to construct a self-adjoint, strictly positive *Transfer Matrix* for Euclidean LGT's with gauge fields and fermions [14]. Both these works are essential reading for every serious student of LGT's. Already in 1977, Creutz [34] had shown how the transfer matrix can be used to relate the Euclidean and Hamiltonian versions of LGT's.

For the purposes of this book it suffices to focus only on the pure-gauge aspects of LGT's. The reason is that the object of primary interest is the flux tube that is expected to be formed between quarks and antiquarks. To avoid going into the details of quark dynamics, which can be quite formidable, one restricts oneself to the case when the quarks and antiquarks are *static*. Therefore all the very deep issues surrounding fermions on the lattice like species doubling, chiral fermions, etc. will not be discussed despite their being very fundamental. For those who wish to pursue those topics in addition to the pure-gauge aspects covered here, the following list of excellent books is highly recommended: *Quarks, Gluons and Lattices* by Creutz [35], *Quantum Fields on a Lattice* by Montvay and Münster [36], *Statistical Field Theory I, II* by Itzykson and Drouffe [37], *Lattice Gauge Theories* by Rothe [38], *Lattice Quantum Field Theory of the Dirac and Gauge Fields* by Baaquie [39], *Quantum Chromodynamics on the Lattice* by Gattringer and Lang [40], *Lattice Gauge Theories and Monte Carlo Simulations* by Rebbi [41], *Discrete Gauge Theory: From Lattices to TQFT* by Oeckel [42], *Introduction to quantum fields on a lattice* by Smit [43], I would also recommend my own set of lectures *Lattice Theory for Non-specialists* [44]. Additionally, the books *An Introduction to the Confinement Problem* by Greensite [45], *Quantum Geometry* by Ambjorn et al. [46] and *Gauge Fields and Strings* by Polyakov [47] are highly recommended.

20.2 An Elementary Introduction to Lattice Field Theories

In this section, we will illustrate all the essential features of lattice QFT's with the help of the most elementary of all relativistic fields, namely, the real scalar field. In the introduction we have briefly discussed its path-integral quantization, as well as its Euclideanization. We explicitly showed that the Euclidean action S_E is positive semi-definite. This is true for almost all QFT's on flat space-time, including gauge theories. The reader is encouraged to check this for Maxwell Electromagnetism, QCD, the Weinberg-Salam electroweak theory, etc. The result follows on noting

$g_{00} = -1$ (Cartesian coordinates) and that all time-like components of vectors pick up factors of i , etc. The situation with fermion fields is more subtle.

Now we proceed to discretize the Euclideanized scalar field theory. We shall simplify our considerations by considering the theory to be in just $D = 1$ space-time; that is, with only temporal dimension. That is of course just quantum mechanics, but we shall pretend it to be a $D = 1$ QFT! The dissatisfied reader can consider more realistic values of D . As far as the essence of LGT's that we are trying to convey, we will see that nothing is compromised by the simplistic choice of $D = 1$. The Euclideanized action is then

$$S_E = \int dx \{(\partial_x \phi)^2 + m^2 \phi^2\} \quad (20.7)$$

Upon discretization x takes only discrete values. Choosing for simplicity a one-dimensional lattice of points equally spaced, we can take these to be $x_n = na$, where a is the lattice spacing. Now the task is to find a substitute for the continuum derivative $\partial_x \phi$. There are clearly many ways of doing this:

$$\partial_x \phi \rightarrow \pm \frac{\phi(x \pm ma) - \phi(x)}{ma}, \frac{\phi(x+a) - 2\phi(x) + \phi(x-a)}{2a} \quad (20.8)$$

and each of these choices leads to inequivalent choices of lattice actions. It is not at all clear that they would all lead to the same continuum theories. We shall come back to this point shortly. Of course, naively they should in the sense that in the limit $a \rightarrow 0$, all these choices coalesce. But in this limit, called the *naive continuum limit*, all the divergences of the theory reappear and one gets back to the ill-defined unregularized QFT's. The continuum limit has to be taken subtly, after due renormalizations, and in lattice field theories, the non-trivial way of reaching the continuum limit is through what is called *the statistical continuum limit* (we will describe the essence of this procedure shortly for our $D = 1$ theory). Then it is a non-trivial issue whether the different choices of lattice actions do indeed describe the same continuum physics!

Let us make the choice $\partial_x \phi(x) = \frac{\phi(x+a) - \phi(x)}{a}$ and proceed. It is an elementary exercise to see that the following *lattice action* S^L emerges from the Euclidean action S_E :

$$S_E \rightarrow S^L = a \sum_n \frac{(\phi_{n+1} - \phi_n)^2}{a^2} + m^2 a \sum_n \phi_n^2 \quad (20.9)$$

where we have adopted the notation $\phi(x_n) \equiv \phi_n$. At this stage, it is customary to scale all dimensionful parameters like fields, masses, coupling constants by appropriate powers of the lattice spacing a so S^L which is dimensionless in the units $\hbar = 1$ is expressed entirely in terms of dimensionless quantities. In $D = 1$, the canonical mass dimension of $\phi(x)$ is $\frac{1}{2}$ (in D space-time dimensions it is $\frac{D-2}{2}$). Therefore, $m^L = m a$, $\phi^L = a^{-\frac{1}{2}} \phi$. The resulting S^L expressed purely in terms of dimensionless quantities is

$$S^L = 2 \sum_n [(\phi_n^L)^2 - \phi_n^L \phi_{n+1}^L] + m_L^2 \sum_n (\phi_n^L)^2 \quad (20.10)$$

Consequently the path-integral, modulo an irrelevant term (potentially singular) of the form $\prod_n a^{\frac{n}{2}}$, takes the form

$$Z = \int \prod_n d\phi_n e^{-(2+m^2) \sum_n \phi_n^2 + 2 \sum_n \phi_n \phi_{n+1}} \quad (20.11)$$

where for the sake of brevity we have suppressed the superscript L on all the quantities.

This is mathematically equivalent to the partition function of a gas of particles in one dimension! Though we have chosen an exceedingly simple case to illustrate this, the demonstration can be extended to most general QFT's in any number of space-time dimensions D . In summary, a D -dimensional QFT is exactly mappable to a problem of *classical* statistical mechanics but in D *spatial* dimensions. This is in so far as all the Wightman functions of the QFT. As briefly mentioned in the introduction, the ramifications of this connection are immense. It allows the powerful techniques of statistical mechanics, both analytical and numerical, to be used to address problems of QFT in a *non-perturbative* way. The reader is referred to Seiler's lecture notes [6] for an excellent exposition of this connection.

20.2.1 The Statistical Continuum Limit

As mentioned before, the lattices in LGT's are ways of regularizing the QFT's in question, in contrast to the physical lattices of, say, the condensed matter systems. In the latter, the lattice structure as well as the lattice spacing(s) have a physical meaning. But when lattices serve the purpose as regularizing QFT's, eventually the lattice spacing has to be taken to zero. But this should be done only after suitable renormalization, just as with cut-offs in the regularization of continuum QFT's.

But the partition function now has no *explicit* dependence on lattice spacing leading to the question as to how then can one take the $a \rightarrow 0$ limit? That leads to one of the most fascinating aspects of LFT's (lattice quantum field theories) called the *statistical continuum limit*. The key to this is the rather elementary observation that as $a \rightarrow 0$, all physical distance scales like Compton wavelengths, physical sizes like charge radii become *infinite* in comparison to the lattice spacing. On the lattice, the physical length scales correspond to *correlation lengths*. Therefore the continuum limit corresponds to the limit in which lattice correlation lengths *diverge*.

So the idea behind the statistical continuum limit of LGT's is to tune various parameters (one of them playing the role of "temperature") of the equivalent statistical system till one or more correlation lengths begin to diverge. But the catch is not all statistical mechanical systems develop diverging correlation lengths. In particular, those that have only *first-order phase transitions* never do. An example of a first-order phase transition is the water-to-ice transition as the temperature of liquid water is lowered to the freezing point. Another example of a first-order phase transition is that of water to steam.

It is useful at this stage to introduce some bare minimum about phase transitions. To keep the discussion manageable, we shall consider only the temperature to be the relevant parameter. The free energy F as a function of $\beta = \frac{1}{kT}$ (T is the temperature and k the Boltzmann constant) codifies the behaviour of the thermodynamic system at equilibrium: a) When $F(\beta)$ is an analytic function for all β , the system has no phase transitions at all (it is in a single phase), b) If $F(\beta)$ is continuous at some β_c but its first derivative (internal energy) is discontinuous, the system undergoes a first-order phase transition at β_c . c) If both $F(\beta)$ and its first derivative are continuous at β_c , but the second derivative diverges, the system is said to have a second-order phase transition at β_c . An example is the so-called λ -point of liquid Helium. Other well-known examples are the Heisenberg Ferromagnet and $d = 2$ Ising model. This is the Landau classification of phase transitions.

There can also be the so-called *infinite order* phase transitions. Such systems are characterized by free energies of the type

$$f(T) = f(T_c) + a e^{-\frac{b}{(T-T_c)^\alpha}} \quad (20.12)$$

with b, α both positive. In other words, the free energy has an essential singularity at T_c . An example of this kind is the Kosterlitz-Thouless transition in the 2-d $x - y$ model [48]. Lattice QCD also possesses a phase transition of this kind with $T_c = 0$, $\alpha = 1$. It is clear that for positive b, α every derivative is continuous at the transition point and yet the free energy is singular; hence the name “infinite order transition”.

An important point worth emphasizing is that at finite volume there can be no phase transitions as the free energy is well defined at each β . Thus only in the so-called *thermodynamic limit* can the free energy develop any non-analytic behaviour. This will have serious implications in practical implementations of lattice field theories.

As far as correlation lengths are concerned, in the thermodynamic limit it diverges for transitions that are second order or higher, including the infinite order transitions. The divergences are characterized by the so-called *critical exponents*. For example, near a second-order phase transition (also called a critical point), the correlation length ξ could diverge as $\xi \simeq |T - T_c|^{-\delta}$, $\delta > 0$. Likewise, near a Kosterlitz-Thouless-type infinite order transition, the divergence could be parametrized as $\xi \simeq e^{b|T-T_c|^{-\eta}}$, $b > 0$, $\eta > 0$. In these cases, δ, η are the critical exponents. It often happens that there are many correlation lengths that diverge the same way. In other words, close to criticality they are simply proportional to each other. In some rare cases, the system may have more than one family of diverging correlation lengths so that within each family the correlation lengths are proportional to each other (close to critical temperature). The other amazing feature of statistical systems is that two systems that are dynamically unrelated can have similarly diverging correlation lengths. For example, the three-dimensional Ising model and the critical point of water! (the reader is referred to Chap. 13 of my book *The Essentials of Thermodynamics* [49] for a discussion of this intriguing connection). This is referred to as *universality*, and phase transitions in systems like 3-d Ising model and water are said to belong to the same *universality class*.

Coming back to the many choices of lattice actions, it is believed that these different choices of latticization lead to systems in the same universality class. Of course this is more of an expectation without any proof. With this crash course on phase transitions, we return to the issue of the statistical continuum limit of lattice regularized field theories. The diverging lattice correlation function(s) could correspond to, for example, various physical masses in units of the lattice spacing, i.e. $\xi = m_{phys} a$. If the correlation function diverges at β_c , $\xi^{-1} = f(\beta)$ such that $f(\beta_c) = 0$. We can then conclude that $a = \frac{f(\beta)}{m_{phys}}$ and as $\beta \rightarrow \beta_c$, $a \rightarrow 0$. Thus on the lattice, the continuum limit is reached by tuning bare parameters appropriately.

Furthermore, a physical quantity of mass dimension d would behave as $A_{phys}^{(d)} \simeq \frac{c \cdot f(\beta)^d}{a^d}$ as $\beta \rightarrow \beta_c$. This is referred to in statistical mechanics as *correlation length scaling*. In particular, dimensionless physical quantities should show no β -dependence in this continuum limit. In practice, that is how one decides whether the continuum limit has been reached or not. Different observables may exhibit differing approaches to the continuum limit. We shall now illustrate analytically how the statistical continuum limit works for our highly elementary $D = 1$ example.

20.2.2 Statistical Continuum Limit of the $D = 1$ Example

Returning to the lattice partition function of Eq. (20.11), we shall calculate the correlation function $\langle \phi_n \phi_{n+N} \rangle$ where $\langle \dots \rangle$ is the usual statistical average, defined by

$$\langle \mathcal{O}(\{\phi_i\}) \rangle = Z^{-1} \int \prod_n d\phi_n \mathcal{O}(\{\phi_i\}) e^{-(2+m^2) \sum_n \phi_n^2 + 2 \sum_n \phi_n \phi_{n+1}} \quad (20.13)$$

To examine the possible existence of phase transitions (the reader is warned that statistical folklore claims the *absence* of any phase transitions in $D = 1$ with finite range interactions), we need to identify the analog of temperature for our system. There are many ways of doing this, but to bring resemblance to the standard models like spin systems, Ising model, etc. we consider Z to be a special case of

$$Z(\beta) = \int \prod_n d\phi_n e^{-(2+m_L^2) \sum_n \phi_n^2 + \beta \sum_n \phi_n \phi_{n+1}} \quad (20.14)$$

with $\beta = 2$. The significance of this value will come out shortly. The β here is related to the temperature as before.

The Transfer Matrix: We now introduce the elegant and highly powerful concept of the *Transfer Matrix* of statistical mechanics to settle the existence of phase transitions in our $D = 1$ model. The work of Lüscher that we alluded to before [14] involved a proof that the Transfer matrix for lattice gauge theories was strictly positive. The transfer matrix for our $D = 1$ statistical system is defined as the *infinite-dimensional*

matrix T whose rows and columns are labelled by the continuous variables ϕ_n, ϕ_{n+1} (with A standing for $2 + m_L^2$):

$$T[\phi_{n+1}, \phi_n] \equiv e^{-\frac{A}{2}(\phi_n^2 + \phi_{n+1}^2) + \beta \phi_n \phi_{n+1}} \quad (20.15)$$

We have chosen a representation of T that is real and symmetric. Therefore T is formally (because of the infinite dimensionality) a Hermitean matrix with real eigenvalues. In the thermodynamic limit $N \rightarrow \infty$, we can replace the real line representing space (actually time here) by a very large circle and by imposing periodic boundary conditions $\phi_N = \phi_1$. Then it follows that $Z(\beta) = \text{tr } T^N$. Eventually N is taken to infinity. Therefore, to determine Z it suffices to determine the largest eigenvalue Λ_0 of T . Various correlation functions can be related to the full spectrum of T . To determine Λ_0 we write down the corresponding eigenvalue equation:

$$\int d\phi_n T[\phi_{n+1}, \phi_n] \Psi_0(\phi_n) = \Lambda_0 \Psi_0(\phi_{n+1}) \quad (20.16)$$

It is straightforward algebra to show that the solution to this is

$$\Psi_0(\phi_n) = \left(\frac{\alpha}{\pi}\right)^{\frac{1}{4}} e^{-\frac{\alpha}{2} \phi_n^2} \quad \Lambda_0 = \sqrt{\frac{2\pi}{A + \alpha}} \quad \alpha = \sqrt{A^2 - \beta^2} \quad (20.17)$$

Likewise, the eigenvalue equation for the first excited state is

$$\int d\phi_n T[\phi_{n+1}, \phi_n] \Psi_1(\phi_n) = \Lambda_1 \Psi_1(\phi_{n+1}) \quad (20.18)$$

with $\langle \Psi_0 | \Psi_1 \rangle = 0$. It is again straightforward to solve for Ψ_1 and Λ_1 :

$$\Psi_1(\phi_n) = \sqrt{2\alpha} \phi_n \Psi_0(\phi_n) \quad \Lambda_1 = \frac{\beta}{A + \alpha} \Lambda_0 \quad (20.19)$$

It is easy to verify that $\Lambda_1 \leq \Lambda_0$. The correlation function of interest to us is $\langle \phi_n \phi_{n+M} \rangle$, and by now standard procedures give for this, for large M ,

$$\langle \phi_n \phi_{n+M} \rangle = c_n c_{n+M} \left(\frac{\Lambda_1}{\Lambda_0}\right)^M \quad (20.20)$$

where c_n, c_{n+M} are some overlap constants (actually independent of M). But this correlation function is also equal to $e^{-\frac{M}{\xi}}$ where ξ is the analog of the spin-spin correlation length. Clearly this correlation length diverges when $\Lambda_1 = \Lambda_0$. Thus our simple $D = 1$ stat mech system does indeed have a diverging correlation length, and the statistical continuum limit of our lattice theory exists. Now we determine the β_c at which this occurs. The condition for this is

$$\beta_c = A + \sqrt{A^2 - \beta_c^2} \rightarrow \beta_c = A \approx 2 \quad (20.21)$$

In the last step we approximated $A = 2 + m_L^2 \approx 2$, as $m_L^2 \rightarrow 0$ at the critical point. This is valid only at the leading order. At sub-leading orders, even the m_L^2 terms have to be taken into account. We end this discussion by determining the “critical exponent” for our model. For this, we determine the behaviour of $\frac{A_1}{A_0}$ in the vicinity of $\beta_c = 2$, i.e. at $\beta = 2 - \delta$ for small and positive δ . It is not difficult to show that $\xi(\beta) \approx (\beta - \beta_c)^{-\frac{1}{2}}$, so the critical exponent is 0.5. There is only one non-trivial correlation length in this example. In [44], the transfer matrix formalism is applied to solve one-dimensional Ising chain and the planar rotator chain.

20.3 Gauge Fields on Lattices

Now we turn to putting gauge theories, both abelian and non-abelian, on lattices. As compared to scalars and Dirac fields (in that order), this is considerably trickier. The main difficulty is discretizing both the actions and gauge transformation laws such that the former are *exactly* invariant under the latter. In the abelian case this is straightforward.

20.3.1 Abelian Gauge Fields

The abelian gauge theory of Maxwell fields (we shall not go into the matter fields either bosonic or fermionic) is described by the action (in the metric $(-1, +1, +1, +1)$ in $D = 4$)

$$S = -\frac{1}{4} \int d^3x dt F_{\mu\nu} F^{\mu\nu} \quad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \quad (20.22)$$

invariant under $A_\mu \rightarrow A_\mu + \partial_\mu \Lambda(x)$, for arbitrary $\Lambda(x)$. It is to be noted that $\Lambda(x)$ is dimensionless. The additional step needed for Euclideanizing this action, over and above what was needed in the scalar case, is that the time-like components A_0 have to be Euclideanized according to $A_0 \rightarrow i A_4$. Then it is a simple matter to show that the Euclideanized action is again positive semi-definite:

$$S_E = \int d^4x F_{\mu\nu} F_{\mu,\nu} \geq 0 \quad (20.23)$$

Let us introduce the so-called forward derivative

$$\Delta_\mu f(\mathbf{x}) \equiv f(\mathbf{x} + a \hat{e}_\mu) - f(\mathbf{x}) \quad (20.24)$$

with a being the lattice spacing and $\hat{e}_\mu$ being the unit vector in the lattice direction μ . We shall assume the lattice to be a $D = 4$ hypercube. With the help of the forward derivative, the field-strength tensor can be discretized as

$$F_{\mu\nu} = \frac{1}{a} (\Delta_\mu A_\nu - \Delta_\nu A_\mu) \quad (20.25)$$

As in the scalar case, all fields and parameters that are dimensionful are scaled by appropriate powers of the lattice spacing to get dimensionless lattice analogs

$$A_\mu^L \equiv a \cdot A_\mu \quad F_{\mu\nu}^L \equiv a^2 F_{\mu\nu} \quad (20.26)$$

One choice for the discrete version of the gauge transformation is

$$A_\mu \rightarrow A_\mu + \frac{1}{a} (\Lambda(\mathbf{x} + a \hat{e}_\mu) - \Lambda(\mathbf{x})) = A_\mu + \frac{1}{a} \Delta_\mu \Lambda(\mathbf{x}) \quad (20.27)$$

In terms of the dimensionless fields, these take the form

$$F_{\mu\nu}^L = \Delta_\mu A_\nu^L - \Delta_\nu A_\mu^L \quad A_\mu^L \rightarrow A_\mu^L + \Lambda(\mathbf{x} + a \hat{e}_\mu) - \Lambda(\mathbf{x}) \quad (20.28)$$

It is easy to check that the discretized field strength is indeed exactly invariant under the discrete version of the gauge transformations. The partition function for this version of QED is

$$Z = \int \prod_{\mathbf{n}, \mu} dA_{\mathbf{n}, \mu}^L e^{-\sum_{\mathbf{n}} (F_{\mu\nu}^L)^2} \quad (20.29)$$

where we have used $\mathbf{n}$ to label the sites of the lattice. Because of the discrete gauge invariance, configurations differing by such gauge transformations make the same contribution to the path-integral and there being infinitely many such configurations, the path integral diverges. This is not the same kind of divergences that are encountered in perturbative QFT's. This sickness exists in the continuum version of the QED path-integral too. There the way out is to use the so-called *gauge fixing* and follow the Faddeev-Popov techniques. Here too the same remedy can be sought. We shall not go into that anymore. Instead, we turn to the problem of putting non-abelian gauge fields on the lattice, which will also point a way to avoid such divergences for QED due to gauge invariance for a different abelian gauge theory based on the compact Lie group $U(1)$.

20.4 Non-Abelian Gauge Fields on Lattice

We now turn to a discussion of putting non-abelian gauge field theories on a lattice. The continuum description of such fields has been discussed at length in Chap. 19 in the case of QCD where the relevant non-abelian group is $SU(3)$. Actually, in that chapter, non-abelian gauge theories based on arbitrary Lie groups were discussed. We simply take over that continuum description to arrive at their lattice formulation.

So, instead of the single gauge field $A_\mu(x)$ of QED, we have now as many gauge fields as the dimension of the non-abelian group G . That is, the gauge fields are A_μ^a with the group index a taking the values $1, 2, \dots, \dim(G)$. As discussed in Chap. 19,

using the Hermitean generators L_a of the group G in its fundamental representation satisfying

$$[L_a, L_b] = i f_{ab}^c L_c \quad \text{Tr}_f(L_a L_b) = \frac{1}{2} \delta_{ab} \quad (20.30)$$

where f_{ab}^c are the structure constants of the group G . One constructs the so-called Lie-algebra-valued vector field $\mathbf{A}_\mu$ according to

$$\mathbf{A}_\mu(x) = L_a A_\mu^a(x) \quad (20.31)$$

In QED, the field strength $F_{\mu\nu}$ was *invariant* under the gauge transformations. In the non-abelian theory, there is no such gauge-invariant tensor. As elaborated in Chap. 19, the non-abelian gauge transformations for the theory with gauge coupling g_0 are

$$\mathbf{A}'_\mu = g(x) \mathbf{A}_\mu g^\dagger(x) - \frac{i}{g_0} g(x) \partial_\mu g^\dagger(x) \quad (20.32)$$

with $g(x)$ being an element of the gauge group locally. The non-abelian field strength (lie algebra valued) is given by

$$\mathbf{F}_{\mu\nu}(x) = \partial_\mu \mathbf{A}_\nu(x) - \partial_\nu \mathbf{A}_\mu(x) + i g_0 [\mathbf{A}_\mu(x), \mathbf{A}_\nu(x)] \quad (20.33)$$

transforming under gauge transformations as

$$\mathbf{F}'_{\mu\nu} = g(x) \mathbf{F}_{\mu\nu} g^\dagger(x) \quad (20.34)$$

Therefore, the field strength unlike in QED is not invariant under gauge transformations. Instead, it transforms covariantly as above. Now we are in a position to discuss the challenges of putting non-abelian gauge theories on the lattice. Let us first consider discretizing the gauge transformations. If we were to follow our earlier approaches in the cases of the scalar fields as also abelian gauge theories, one may be tempted to consider

$$\mathbf{A}'_\mu(x) = g(x) \mathbf{A}_\mu(x) g^\dagger(x) - \frac{i}{g_0} g(x) \frac{\{g^\dagger(x + a \hat{e}_\mu) - g^\dagger(x)\}}{a} \quad (20.35)$$

A tedious but straightforward calculation reveals that under this discrete form of gauge transformations, the discretized field strength

$$\mathbf{F}_{\mu\nu} = \Delta_\mu \mathbf{A}_\nu(x) - \Delta_\nu \mathbf{A}_\mu(x) + i g_0 [\mathbf{A}_\mu(x), \mathbf{A}_\nu(x)] \quad (20.36)$$

does not transform according to Eq. (20.34). This is not due to the shortcomings of the particular discretization of the field strength and the gauge transformation; it is a rather deep feature. There simply is no discretization that works!

The resolution actually rests on some fundamental mathematical properties of non-abelian gauge fields. We shall simply show how it works. The basic object to

be considered is the so-called *link variable*. It is a special case of what are called holonomy elements. It is a *bilocal* object defined as

$$U_{\mathbf{x},\mathbf{y}} \equiv P e^{i g_0 \int_{\mathbf{x}}^{\mathbf{y}} \mathbf{A}_{\mu} dx^{\mu}} \quad (20.37)$$

The integration is a line integral along some curve Γ . U is also called a *path-ordered exponential* and its precise meaning is obtained by dividing the curve into a very large number of very small segments $x x_1, x_1 x_2, \dots, x_n y$ and defining the path-ordered exponential as

$$P e^{\int_{\mathbf{x}}^{\mathbf{y}}} \equiv (e^{\int_{\mathbf{x}}^{\mathbf{x}_1}})(e^{\int_{\mathbf{x}_1}^{\mathbf{x}_2}}) \dots (e^{\int_{\mathbf{x}_n}^{\mathbf{y}}}) \quad (20.38)$$

in the limit $n \rightarrow \infty$. Now we prove a remarkable property of path-ordered exponential in Eq. (20.37). For that, let us consider how the tiny segments transform under non-abelian gauge transformations (where $x' = x + \Delta x$):

$$\begin{aligned} e^{i g_0 \int_{\mathbf{x}}^{\mathbf{x}+\Delta\mathbf{x}} dx^{\mu} \mathbf{A}'_{\mu}(x)} &\approx 1 + i g_0 \Delta x^{\mu} \mathbf{A}'_{\mu}(x) \\ &= 1 + i g_0 \Delta x^{\mu} g(x) \mathbf{A}_{\mu}(x) g^{\dagger}(x') - \frac{i}{g_0} \{g(x) \frac{(g^{\dagger}(x') - g^{\dagger}(x))}{\Delta x^{\mu}}\} \\ &= i g_0 \Delta x^{\mu} g(x) \mathbf{A}_{\mu} g^{\dagger}(x') + g(x) g^{\dagger}(x') \\ &= g(x) e^{i g_0 \int_{\mathbf{x}}^{\mathbf{x}'} dx^{\mu} \mathbf{A}_{\mu}(x)} g^{\dagger}(x') \end{aligned} \quad (20.39)$$

With the help of this and the unitarity of $g(x)$, i.e. $g(x)g^{\dagger}(x) = 1$, it immediately follows that under gauge transformations

$$U'_{\mathbf{x},\mathbf{y}} = g(x) U_{\mathbf{x},\mathbf{y}} g^{\dagger}(y) \quad (20.40)$$

This extremely elegant result forms the central basis for LGT's. For that reason, we have carefully worked out the intermediate steps leading to it. When the path in question is a link on the lattice, i.e. what joins two adjacent sites, the path-ordered exponential becomes the so-called *Link variable* and all quantities of interest for LGT's are built out of them.

20.4.1 Invariants of Non-Abelian Gauge Theories

Now we come to the crucial issue of objects that are invariant under the non-abelian gauge transformations. Among other things, candidates for the action for non-abelian gauge theories, which has to be necessarily gauge-invariant, will also be determined by this analysis. Let us first address this in the continuum version and then take up the problem on lattices.

As already noted, the field strength is not gauge-invariant. Instead it transforms homogeneously (as opposed to the gauge field $\mathbf{A}_{\mu}(x)$ which transforms inhomogeneously). In mathematical parlance, $\mathbf{A}_{\mu}$ is akin to connections, while $\mathbf{F}_{\mu\nu}$ akin to

curvature. There are very beautiful geometric and topological descriptions of gauge theories which we shall not go into. Not only the field strengths, but even Lorentz-invariants constructed out of them like $\mathbf{F}_{\mu\nu}\mathbf{F}^{\mu\nu}$, $\mathbf{F}_{\mu\nu}\mathbf{F}^{\nu\sigma}\mathbf{F}^{\mu}_{\sigma}$ are not gauge-invariant. In fact all of them transform in the same homogeneous way as $\mathbf{F}_{\mu\nu}$ itself! However, all these are still Lie-algebra valued, and, in particular, are matrices. Therefore, taking their traces or determinants would indeed give gauge-invariant objects. The simplest of them is $\text{Tr}(\mathbf{F}_{\mu\nu}\mathbf{F}^{\mu\nu})$ and that gives the candidate Lagrangean density

$$\mathcal{L} = -\frac{1}{2} \text{Tr} \mathbf{F}_{\mu\nu}\mathbf{F}^{\mu\nu} = -\frac{1}{4} F_{\mu\nu}^a F^{\mu\nu a} \quad (20.41)$$

This is the non-abelian analog of the Lagrangean density for QED. Though they look similar, there are profound differences between them. In the non-abelian case $F_{\mu\nu}^a$ has terms both linear and quadratic in A_{μ}^a . Thus the non-abelian gauge theory is a self-interacting theory, and these self-interactions produce dramatic physical differences. We now address the issue of gauge-invariant objects on the lattice.

20.4.2 Wegner-Wilson Loops

The gauge transformation properties of link variables are also what Wegner first found in [1], and subsequently by Jan Smit, Polyakov and Wilson. The simple transformation law for link variables has deep ramifications. To appreciate them, consider a closed-loop W (for both Wilson and Wegner!) on a hypercubic lattice (generalization to other types of lattices is straightforward). The segments of the loop are the links of the lattice. It is not necessary for the loop to be *Planar*. Now the Wegner-Wilson loop variable is constructed by taking the product of the link variables $U_{\mathbf{x},\mathbf{y}}$ taken in the same order. Such loop variables had already been constructed and studied by Wegner. He had even introduced the so-called *dual* loop variables, later rediscovered by Gerard t'Hooft, and are called t'Hooft loops. Wilson in his seminal work [5] rediscovered these loop variables. So, what is special about these Wilson loops? To answer that, we work out how they transform under non-abelian gauge transformations. We first give the definition of these loops:

$$\mathbf{W} \equiv U_{x_1, x_2} U_{x_2, x_3} \dots U_{x_n, x_1} \quad (20.42)$$

where $x_1, x_2, \dots, x_n$ are the ordered vertices making up the loop. Under the non-abelian gauge transformations, the link variables transform as $U_{x,y} \rightarrow g(x) U_{x,y} g^{\dagger}(y)$. Consequently, $\mathbf{W}$ transforms as (on using the shorthand notation $U_{x_i, x_j} \equiv U_{i,j}$)

$$\begin{aligned} \mathbf{W}' &= g(x_1) U_{1,2} g(x_2)^{\dagger} g(x_2) U_{2,3} g(x_3)^{\dagger} \dots g^{\dagger}(x_n) g(x_n) U_{n,1} g^{\dagger}(x_1) \\ &= g(x_1) \mathbf{W} g^{\dagger}(x_1) \end{aligned} \quad (20.43)$$

Therefore $\mathbf{W}$ transforms exactly as the continuum field strength. There, since $\text{Tr} \mathbf{F}_{\mu\nu} = 0$, the first non-trivial invariant had to be quadratic in the field strength.

But for the Wilson loop, there is no reason for its trace to vanish, so its trace is the simplest gauge-invariant that can be constructed, i.e.

$$\text{Tr } \mathbf{W}' = \text{Tr } g(x) \mathbf{W} g^\dagger(x) = \text{Tr } \mathbf{W} \quad (20.44)$$

where the cyclic property of traces was used (justified since all the matrices involved are finite-dimensional), along with $g(x)g^\dagger(x) = 1$. Therefore, gauge-invariants on lattice can be obtained by considering arbitrary loops, non-planar, and self-intersecting ones included, and computing the trace of the corresponding Wilson loop. This happens in the continuum too, where traces of arbitrary Lorentz-invariant products of field strengths and their gauge-covariant derivatives yield invariants. It is just that the construction on the lattice is much simpler, and intrinsically geometric. In fact traces of all powers $\mathbf{W}^n$ are also gauge-invariant. Geometrically, such powers can always be associated with a single loop that is tantamount to going around the original loop several times.

20.4.3 Polyakov Lines (Loops)

Often periodic boundary conditions are used on lattices. In a Euclidean lattice, there is no intrinsic difference between spatial and temporal directions. Any direction can be arbitrarily treated as temporal. Periodic boundary conditions along such a temporal direction lead to yet another class of gauge-invariant objects called *Polyakov Loops* which are products of link variables along the temporal direction. Because of the periodic boundary conditions, the end-points get identified and in effect one has a closed loop. Consequently, traces of Polyakov loops are also gauge-invariant. We shall extensively use Polyakov Loop correlation functions as efficient tools to study Yang-Mills flux tubes equivalently the static $Q\bar{Q}$ -potentials. Wilson loops are also capable of measuring the same static potential. We shall elaborate this connection in the next chapter.

20.4.4 The Plaquette Action

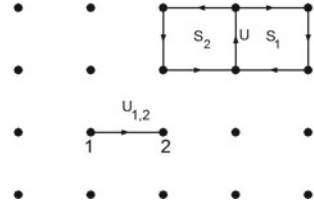
The simplest Wilson loop on a hypercubic lattice is a square bounded by links in two distinct directions, say, μ and ν . It is called a *Plaquette*, a French word meaning a square. The plaquette variable $P_{\mathbf{x},\mu\nu}$ is defined in terms of the four link variables spanning it by

$$P_{\mathbf{x},\mu\nu} \equiv U_{\mathbf{x},\mu} U_{\mathbf{x}+\hat{\mu},\nu} U_{\mathbf{x}+\hat{\nu},\mu}^\dagger U_{\mathbf{x},\nu}^\dagger \quad (20.45)$$

Note that a link variable in the backward direction on the lattice is to be taken as the inverse of the link variable on the same link. More precisely, $U_{\mathbf{x}+a\hat{\nu},-\nu} = U_{\mathbf{x},-\nu}^\dagger$. We will examine the trace of the plaquette variable, which is gauge invariant, when the lattice spacing is very small. Then,

$$U_{\mathbf{x},\mu} \approx 1 + i g_0 a \mathbf{A}_\mu(x) + \dots \quad (20.46)$$

Fig. 20.1 Important elements for lattice gauge theories



In matrix equations, 1 should be understood as the unit matrix of appropriate dimensions. A tedious, but straightforward calculation yields

$$\text{Tr} (1 - P_{\mathbf{x},\mu\nu}) \approx \frac{g_0^2}{2} a^4 \text{Tr} \mathbf{F}_{\mu\nu} \mathbf{F}_{\mu\nu} + \dots \quad (20.47)$$

where the dots indicate terms of higher order in lattice spacing. No summation convention is invoked. Summing this over all the pairs of directions μ, ν , one gets

$$\frac{2}{g_0^2} \sum_{\mathbf{x}, \mu, \nu} \text{Tr} (1 - P_{\mathbf{x},\mu\nu}) = \sum_{\mathbf{x}} a^4 \text{Tr} \mathbf{F}_{\mu\nu} \mathbf{F}_{\mu\nu} \rightarrow \int d^4x \text{Tr} \mathbf{F}_{\mu\nu} \mathbf{F}_{\mu\nu} \quad (20.48)$$

which is the continuum action. We are still in Euclidean space.

This suggests the candidate action for lattice gauge theories as

$$S_L = \frac{4}{g_0^2} \sum_P \text{Tr} (1 - P) \quad (20.49)$$

where the summation is over all the plaquettes. This is manifestly positive semi-definite because each P is a unitary matrix. It should however be stressed that trace of any combination of Wilson loops can also be a candidate action as long as it satisfies the Osterwalder-Seiler positivity condition.

In Fig. 20.1 we display the important elements for lattice gauge theories like link variables $U_{1,2}$, plaquettes, one formed by U and S_1 , the other by U and S_2 , etc. The open products of links S_1, S_2 are also called *staples*.

20.4.5 The LGT Path-Integral

We are almost ready to write down the path-integral for LGT's. The integration is over all the link variables which are elements of the gauge group G . What is left is choosing an appropriate measure for the path-integral. Since the integration is over group elements, there is a natural measure from group theory called the Haar measure dU . For compact gauge groups like $SU(2)$, $SU(3)$, the Haar measure can be normalized such that $\int dU = 1$. The Haar measure is also *group invariant* which means that for fixed elements g, h of the group, $d(gU) = d(Uh) = dU$. As under a gauge transformation $U'_{\mathbf{x},\mu} = g(\mathbf{x})U_{\mathbf{x},\mu}g^\dagger(\mathbf{x}')$, it follows that the Haar measure is also gauge-invariant. Thus lattice regularization is manifestly gauge-invariant and

the path-integral exists without the need for any *gauge fixing*! Writing down the path-integral for LGT's in the purely gauge-field sector (that is the only sector we shall be dealing with in the rest of the book),

$$Z(\beta)_{LGT} = \int \prod_l dU_l e^{-\beta \sum_p \text{Tr}(1 - P)} \quad (20.50)$$

where $\beta = \frac{4}{g_0^2}$. In the light of our earlier remarks, the bare coupling $g_0(a)$ has to be tuned till one finds a phase transition with a diverging correlation length. It turns out that this happens as $g_0 \rightarrow 0$. Despite this happening at vanishing bare coupling constant, the LGT's give a non-perturbative treatment of the theory at very strong physical coupling constant! In Wilson's pioneering paper [5] he proved several interesting results but in the limit of strong bare coupling, i.e. $g_0 \rightarrow \infty$, but unfortunately that's not where the statistical continuum limit resides. The challenge is to prove properties like confinement, spontaneous breaking of chiral symmetry, etc. in the regime of weak bare coupling. This will be the main theme of the remaining chapters.

However, there are advantages of gauge fixing as, for example, in setting up a transfer matrix. It is possible to gauge-fix such that all links in a particular direction, which can be taken to be the temporal direction, can be fixed to be unity, unless there are periodic boundary conditions along that direction, in which case all but one of the links in that direction can be fixed to be unity.

This ends our introduction to lattice gauge theories to the extent needed to appreciate the developments discussed in the remaining chapters of the book. The reader is referred to [13, 14] for many important properties of LGT's like positivity, etc.

References

1. F. Wegner, J. Math. Phys. **12**, 2259 (1971)
2. H.A. Kramers, G.H. Wannier, Phys. Rev. **60**, 252 (1941)
3. J. Kogut, L. Susskind, Phys. Rev. D **11**, 395 (1975)
4. J. Kogut, Rev. Mod. Phys. **51**, 659 (1979)
5. K.G. Wilson, Phys. Rev. D **10**, 2445 (1974)
6. E. Seiler, *Gauge Theories as a Problem of Constructive Quantum Field Theory and Statistical Mechanics*. Lecture Notes in Physics, vol. 159 (Springer, 1982)
7. G. Mack, V.B. Petkova, Ann. Phys. NY **23**, 442 (1979)
8. J. Fröhlich, Phys. Lett. B **83**, 195 (1979)
9. F.A. Berezin, *The Method of Second Quantization* (Academic, New York, 1966)
10. J. Smit, unpublished, 1972–73. Private communication to K.G. Wilson from R.J. Finkelstein, with unpublished documentation of Jan Smit's studies of lattice gauge theory while he was a graduate student at UCLA
11. A.M. Polyakov unpublished; see also Phys. Lett. B **59**, 82 (1975)
12. K.G. Wilson in *Future Directions in Particle Theory*, 1983 Lepton-Photon Symposium, Cornell
13. K. Osterwalder, E. Seiler, Ann. Phys. **110**, 440 (1978)
14. M. Lüscher, Comm. Math. Phys. **54**, 283 (1977)
15. R.P. Feynman, Rev. Mod. Phys. **20**, 267 (1948)
16. R.P. Feynman, A.R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw Hill, 2010)

17. S. Weinberg, *The Quantum Theory of Fields - I* (Cambridge University Press, 1984)
18. N.D. Hari Dass, *Dirac and the Path Integral* (2020) [arXiv:2003.12683](https://arxiv.org/abs/2003.12683) [hist-ph]
19. H. Kleinert, *Path Integrals* (World Scientific, 2009)
20. B.E. Baaquie, *Quantum Finance* (Cambridge University Press)
21. A.M. Polyakov, Phys. Lett. B **103**, 207 (1981)
22. J. Polchinski, *String Theory - I* (Cambridge University Press, 2013)
23. J. Schwinger, Proc. Nat. Acad. Sc. **44** 956 (1958)
24. T. Nakano, Phys. Rev. **115**, 721 (1959)
25. J. Schwinger, *Four-Dimensional Euclidean Formulation of QFT*, in ICHEP58
26. J. Schwinger, *Particles, Sources, and Fields - I* (Addison-Wesley Publishing Company, 1970)
27. R. Haag, *Local Quantum Physics*, 2nd edn. (Springer Publications, 1991)
28. K. Osterwalder, R. Schrader, Comm. Math. Phys. **31**, 83 (1973)
29. K. Osterwalder, R. Schrader, Comm. Math. Phys. **42**, 281 (1975)
30. B. de Wit, J. Smith, *Field Theory in Particle Physics - I* (North Holland Personal Library, 2012)
31. S.W. Hawking, W. Israel, (eds.), *General Relativity - An Einstein Centenary Survey* (Cambridge University Press, 2010)
32. M. Creutz, Phys. Rev. D **21**, 2308 (1980)
33. N.D. Hari Dass, Int. J. Mod. Phys. B **14** 1989 (2000)
34. M. Creutz, Phys. Rev. D **15**, 1128 (1977)
35. M. Creutz, *Quarks, Gluons and Lattices*. Cambridge Monographs in Mathematical Physics (Cambridge University Press, 1983)
36. I. Montvay, G. Munster, *Quantum Fields on a Lattice* (Cambridge University Press, 1994)
37. C. Itzykson, J.M. Drouffe, *Statistical Field Theory*, Vols. I and II. Cambridge Monographs on Mathematical Physics (Cambridge University Press, 1988)
38. H.J. Rothe, *Lattice Gauge Theories*. Lecture Notes on Physics (World Scientific, 2012)
39. B.E. Baaquie, *Lattice Quantum Field Theory of the Dirac and Gauge Fields* (World Scientific, 2020)
40. C. Gattringer, C. Lang, *Quantum Chromodynamics on the Lattice* (Springer, 2009)
41. C. Rebbi, *Lattice Gauge Theories and Monte Carlo Simulations* (World Scientific, 1983)
42. R. Oeckel, *Discrete Gauge Theory: From Lattices to TQFT* (Imperial College Press, 2005)
43. J. Smit, Introduction to quantum fields on a lattice. Camb. Lect. Notes Phys. **15**, 1–271 (2002)
44. N.D. Hari Dass, *Lattice Theory for Non-specialists*, NIKHEF-H 84/11 Oct 1984
45. J. Greensite, *An Introduction to The Confinement Problem*. Lecture Notes on Physics, vol. 821 (Springer, 2011)
46. J. Ambjorn, B. Durhuus, T. Jonsson, *Quantum Geometry*. Cambridge Monographs on Mathematical Physics (Cambridge University Press, 2021)
47. A.M. Polyakov, *Gauge Fields and Strings* (Harwood Academic Publishers, 1987)
48. J.M. Kosterlitz, G.J. Thouless, J. Phys. **C6**, 1181 (1973)
49. N.D. Hari Dass, *Essentials of Thermodynamics* (SRI Books, Singapore, 2020)

Part IV

Strings Regained: From Yang-Mills Flux Tubes to Effective String Theories

Lattice Gauge Theory and Yang-Mills Flux Tubes

21

21.1 Introduction

In Sect. 8 of our Chap. 19 on QCD, we gave a qualitative introduction to one of the most challenging problems for QCD, namely, that of Colour Confinement, and in particular that of Quark Confinement. This was necessitated by the observational fact that quarks have never been seen as free asymptotic states even during the most energetic of hadronic collisions. Though in two dimensions, it had been shown by Schwinger [1] that electrons are absent as asymptotic states in $D = 2$ QED, and still manifest themselves in deep inelastic phenomena (see Wilson's pioneering paper on quark confinement that introduced lattice gauge theories [2] for some interesting discussions on this), generalization of such a mechanism to four dimensions has not been realized even to this day.

Furthermore, the general view evolved in favour of this confinement to be *permanent* or *absolute*, in the sense that even arbitrarily high energies would not liberate quarks from hadrons. Of course, the big question was whether QCD or, for that matter, any QFT could really achieve such a miracle! As already elaborated in that section, it was expected on heuristic grounds that a theory like QCD with *Asymptotic Freedom*, which caused renormalized couplings to *decrease* with increasing scale, would also cause coupling to grow in the infrared (long distances) causing such a permanent confinement. But there was no proof of such a possibility. Looked differently, such a heuristic expectation would require an *absence* of phase transitions separating the strong- and weak-coupling (renormalized) regimes. Such a scenario was indeed contemplated by Migdal [3] (he was also a pioneer in recognizing parallels between QCD and String Theory) almost immediately after Kenneth Wilson's work on quark confinement and lattice gauge theories [2].

A strikingly beautiful suggestion as to the feasibility of realizing a confinement mechanism in physically reasonable theories came in the form of *The Dual Superconductor Mechanism*, also discussed qualitatively in the same section.

Dual superconductivity was first proposed by 'tHooft and Mandelstam [4,5]. Other key references, apart from that of Wilson [2], are [6–12]. A number of technical issues connected with the dual superconductor mechanism have been discussed by Ripka in his review [13]. In this chapter we shall discuss the Yang-Mills flux tube in great detail, based on the pioneering Monte Carlo studies by Creutz [14,15], Ambjorn et al. [16], Bali et al. [17,18], Haymaker et al. [19,20], culminating in recent high accuracy studies by Lüscher and Weisz [21,22], as well as by Pushan Majumdar and myself [23–26]. Other recent investigations of this type are by Kuti and his collaborators [27,28], and by Casselle and collaborators [29]. The excited states of the flux tube have been investigated by Brandt and Majumdar [30,31]. Brandt has revisited these issues in [32] which we shall come back to in the last chapter. A comprehensive survey can be found in my article [33]. While all these studies were based on Monte Carlo numerical simulations, there have been a few attempts to understand the flux tube analytically also. Adler had constructed a dielectric model for them [34], while Baker et al. [35–37] approached the problem from the perspectives of a gauge theory dual to QCD, and exploiting the weakness of the coupling in the dual theory when the original theory (QCD) is in the strong-coupling regime. But it is fair to say that these analytical approaches are no way near the Monte Carlo studies in terms of the precision with which the flux tubes have been understood. Nevertheless, it is important to go back to the analytical approaches and see how far one can improve them in the light of the numerical studies.

In fact, the first analytical approach to the confinement problem was in Wilson's own work of 1974. He did demonstrate confinement, and a picture of flux tubes as essentially the links of the lattice connecting the quark-antiquark pair. But this was in the so-called *strong-coupling* regime, while the statistical continuum limit of the LGT will be seen to be in the extreme weak-coupling limit (see [15]; we will also explain this point shortly). Nevertheless, confinement could persist all the way to the extreme weak coupling limit if there is no second or higher transition separating the two regions. This was the scenario conjectured by Migdal [3]. This situation should be contrasted with the case of compact QED, with gauge group $U(1)$ in $D = 4$ where such a phase transition does indeed separate the confining phase at strong coupling from the coulomb phase at weak couplings.

There are two distinct aspects of the flux tubes that are worthy of serious study. One of them is the *static potential* between a quark-antiquark pair. The dual superconducting mechanism would predict a linearly rising potential (or at least the dominant part of it), while any potential rising with distance would confine the quarks. Thus, the very confirmation of a linearly rising part in the static potential would already be a partial vindication of the dual superconductor picture. We already discussed at length the static potential that would obtain in the perturbative limit of QCD in Sect. 7 of Chap. 19. This will provide a powerful check on the numerical simulations. The other aspect is a direct determination of the profile of the flux tubes. This would involve measuring the chromoelectric and magnetic fields in the presence of the quark-antiquark pair. From these fields, one can work out the energy density and Euclidean action density also. At points very close to either the quark or the antiquark, one would expect a coulombic behaviour, with of course contributions coming from

eight gluonic fields in place of the single photon field of QED. Furthermore, at points slightly away one would expect corrections dictated by the asymptotic freedom of QCD.

We now turn to a case-by-case discussion of the numerical works above. Before that we discuss the important topic of the precise observables of LGT's that would probe both these aspects of the flux tubes. Our discussion will be brief, focussing on the essentials. For these and other details, the reader is encouraged to consult the list of books and references outlined in Chap. 20 on LGT's.

21.2 Flux Tube Observables

21.2.1 The Static $Q\bar{Q}$ -Potential

To see which lattice observable would measure the static $Q\bar{Q}$ -potential, let us consider a static (infinitely heavy) $Q\bar{Q}$ pair separated by a distance R , and let it propagate for Euclidean time interval T . On the one hand, this will increase the Euclidean action by $V(R)T$. On the other hand, to create a gauge-invariant excitation of such a $Q\bar{Q}$ -pair, one has to use $P e^{ig_0 \int_x^y A \cdot dx}$ where x, y are the locations of the quark and antiquark. The world lines of the quark and antiquark would each add a contribution of the type $\int d^4x j_\mu A^\mu$ to the action; using the current for a point particle this would just be $g_0 \int A_\mu dx^\mu(\tau)$, where τ parametrizes the world line. The net effect is the evaluation of the expectation value of a rectangular Wilson loop of size $R \times T$.

This is even more straightforward to show on a lattice. For that, choose a gauge so that all links in the temporal direction of a hypercubic lattice are fixed to be unity (see Chap. 20 for how this can be done). Then using the *Transfer Matrix* formalism (also introduced and applied in Chap. 20), and essentially following the steps in Sect. 20.2.2 there, it is straightforward to show that (see also [19])

$$\langle W_{T \times R} \rangle = \sum_i c_i e^{-T E_i(R)^{q\bar{q}}} \quad (21.1)$$

Here c_i are some overlap constants, exactly as in Eq. (20.20). The ratios $\frac{\Lambda_i}{\Lambda_0}$ of the eigenvalues of the transfer matrix now become $e^{-E_i(R)^{q\bar{q}}}$, with $E_i(R)^{q\bar{q}}$ the energy spectrum in the presence of the quark-antiquark pair separated by R , which can also be interpreted as the energy spectrum of the flux tube. The lowest of these is just the static potential $V(R)$. It is important to note that the above “spectral representation” is valid for all T , not necessarily for very large T . This simple formula exposes a central dilemma in the use of rectangular Wilson loops for determining the static potential; if T is small, or at best moderately large, the contributions of the excited states to $\langle W \rangle$ cannot be neglected, and become a major source of contamination of the static potential. On the other hand, increasing T , while reducing this contamination, also exponentially suppresses the signal. A trade-off, dictated by the available numerical resources, has to be made often.

21.2.2 Confinement Criterion

Let us assume that indeed the vacuum contribution dominates, and $\langle W \rangle \simeq c e^{-TV(R)}$. If there is confinement due to a linearly rising $V(R)$, as in the dual superconductor models, $V(R) = \sigma R$ (σ is referred to as the *string tension*), then one would have $\langle W \rangle \simeq e^{-\sigma RT} = e^{-\sigma A}$ where A is the area of the rectangular loop. This is called the *area law* for confinement. If the loop is not a planar rectangular loop, A will be replaced by the minimal area enclosed by the loop.

If on the other hand, there is no confinement, the potential $V(R)$ falls off with increasing R and the energy of the system, being the self-energies of the quarks, is a constant leading to the behaviour $\langle W \rangle \simeq e^{-cT}$. However, on an Euclidean lattice, there should be a symmetry between R and T , and one should really expect $\langle W \rangle \simeq e^{-c(R+T)} = e^{-cP}$ where P is the perimeter. Thus the Wilson loop expectation value (for large enough loops) can distinguish between linearly rising confinement versus no confinement. Since self-energies are always present, the actual behaviour of Wilson loops are more like $\langle W \rangle \simeq e^{-\sigma A - cP}$.

This confinement criterion works mostly for the pure-gauge sector. When there are *dynamical quarks* in addition, the situation is much more complicated and no simple criterion for quark confinement, such as the area law above, can be put forward. This has to do both with the fact that vacuum polarization effects giving rise to light quark-antiquark pairs can break the flux tube into smaller segments and the related phenomena of quarks and antiquarks combining to form colour neutral hadrons. For further discussions, the reader is referred to Sect. I.3 of [38].

Another distortion of the area law comes from the fact that the “corners” of the Wilson loop (four in the case of rectangular Wilson loops) make Wilson loops somewhat singular from a continuum point. Typically they contribute R -independent but β -dependent terms to the energies, which violate the scaling behaviour required for the statistical continuum limit. So, they contribute to the perimeter part. Creutz suggested a clever way of overcoming both of these problems to a large extent. His proposal was to study the so-called *Creutz Ratios* (he did not name them so, others did!) instead of the Wilson loops themselves. These ratios are such that

$$r_{12;34} \equiv \frac{\langle W_1 \rangle \langle W_2 \rangle}{\langle W_3 \rangle \langle W_4 \rangle} \quad (21.2)$$

where the four loops are so chosen that $P_1 + P_2 = P_3 + P_4$, where the P 's are the perimeters, and $C_1 + C_2 = C_3 + C_4$ with the C 's being the number of corners in each Wilson loop. If all the loops are rectangular, this condition is automatically fulfilled.

21.2.3 The Polyakov Lines (Loops)

It is often customary to choose periodic boundary conditions along one or more directions of the lattice. Designating one such direction to be the temporal direction, a Polyakov Line (Loop) is defined as the ordered product of all the links in the

temporal direction. Trace of this product (also called Polyakov loop, for example, by Lüscher and Weisz [22]) is gauge-invariant because of the periodic boundary conditions. Clearly, at each point of space $\mathbf{x}$ there is such a gauge-invariant Polyakov Loop denoted by $P_{\mathbf{x}}$. Thus (see Chap. 20 for the definition of the link variable $U_{\mathbf{x},\mu}$ at the *space-time* point $\mathbf{x}$ along the positive direction μ):

$$P_{\mathbf{x}} \equiv \text{Tr } U_{\mathbf{x},t} U_{\mathbf{x}+\mathbf{a}\mathbf{e}_t,t} \dots U_{\mathbf{x}+(\mathbf{T}-\mathbf{a}\mathbf{e}_t),t} \quad (21.3)$$

where T is the total extent along the temporal direction. What is of interest to us is the lattice averaged *Polyakov Loop Correlator* $\langle P_{\mathbf{x}}^* P_{\mathbf{x}+\mathbf{R}} \rangle$ of two loops located $\mathbf{R}$ apart. Once again, this average can be worked out using the transfer matrix formalism, just as in the case of the Wilson loop average. The result is the spectral decomposition

$$\langle P_{\mathbf{x}}^* P_{\mathbf{x}+\mathbf{R}} \rangle = \sum_{n=0}^{\infty} w_n e^{-E_n^{q\bar{q}} T} \quad (21.4)$$

Remarkably, the coefficients w_n are *integral*. To understand why this is so, and for a very lucid exposition of Polyakov Loop correlators, the reader is referred to [22]. The purely discrete spectrum of the transfer matrix in finite volume and the non-degenerate nature of the ground state are the essence.

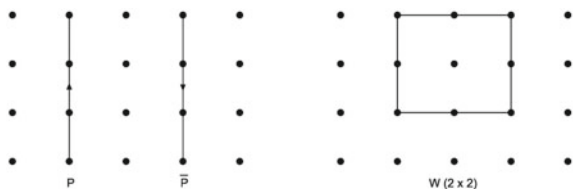
Since T is the largest temporal difference possible on the lattice, it follows from this spectral decomposition that the Polyakov Loop correlator achieves the best projection onto the ground state albeit at the expense of much weaker signals. The criteria for confinement, or the lack of it, are exactly the same as in the case of the Wilson loops. However, there is no need to form the analogs of the Creutz Ratios.

In the figure below we show Wilson loops as well as Polyakov correlators on a lattice (Fig. 21.1).

21.2.4 The Flux Tube Profile

The other important aspect of flux tubes is their *profile*, though this aspect is also intimately tied up with the static potential aspect. Broadly speaking, a profile would be how a flux tube “looks” in space. As they are manifestation of the gauge fields in the presence of a quark-antiquark pair, one would need infinitely many gauge-invariant observables to characterize the profiles of flux tubes. Even if one were to restrict this to half a dozen or so, visualizing the flux tube in terms of them would still be intractable. In their important studies of this issue, Bali et al. on the one

Fig. 21.1 Wilson loops and Polyakov loop correlators



hand and Haymaker et al. on the other [17,20] sought to essentially measure the chromoelectric and chromomagnetic fields in the vicinity of a large Wilson loop. But as already explained in our Chap. 19 on QCD, field strengths of a non-abelian gauge theory like QCD (in contrast to the field strengths of QED) are not gauge-invariant. In fact, even the concept of a flux is not gauge-invariant. So what these authors settled down for is a study of the *energy density* and *action density* both of which are gauge-invariant. Even though in the continuum these quantities are local, the lattice observables that are equivalent to them are still pretty intricate. Actually in the classical continuum limit all of $Tr \{F_{\mu\nu} F_{\alpha\beta}\}$ are gauge-invariant for all $(\mu, \nu, \alpha, \beta)$. But even these are intractable as far as visualizing the flux tube in concerned. Haymaker et al. use the simpler subset $Tr \{F_{\mu\nu} F_{\mu\nu}\}$ where no *summation convention* has been adopted and there are six such gauge-invariants. Even if one is not close enough to the continuum limit where rotational invariance has been restored, the subgroup of rotation group that maps the hypercubic lattice to itself (for example, in two dimensions these would be rotations through 90°) is enough to consider the combinations that would correspond in the continuum to E^2, B^2 . They suffice to yield, again in the continuum the energy and action densities.

Both these groups studied the connected correlation function between a Wilson loop and a plaquette. The uses of these correlation functions as suitable observables for the study of flux tube profiles had been suggested much earlier by several people [39]. We give below the formulae in [20]. Those used by [17] are closely related:

$$f_{\mu\nu}(x)^L \equiv \beta \frac{\langle W Tr P_{\mathbf{x},\mu\nu} \rangle - \langle W \rangle \langle Tr P_{\mathbf{x},\mu\nu} \rangle}{\langle W \rangle} \quad (21.5)$$

where W is a suitably large Wilson loop, and $P_{\mathbf{x},\mu\nu}$ a plaquette (see Chap. 20 on LGT's) located at $\mathbf{x}$ in the μ, ν plane. In the naive continuum limit, this approaches [20]

$$f_{\mu\nu}^L(x) \rightarrow -a^4 \langle Tr F_{\mu\nu} F_{\mu\nu} \rangle_{q\bar{q}} \quad (21.6)$$

again, no summation convention has been used. Though this correspondence holds strictly only in the naive continuum limit $a \rightarrow 0$ both groups use the connected Wilson loop correlation function as a measure of the continuum quantities. Therefore, at finite lattice spacing, or equivalently away from the continuum limit, what are measured are not strictly the action and energy densities. In terms of the chromoelectric and chromomagnetic components [20]

$$f_{\mu\nu}(x) \equiv \frac{f_{\mu\nu}^L(x)}{a^4} \rightarrow \frac{1}{2}(-B_1^2, -B_2^2, -B_3^2, E_1^2, E_2^2, E_3^2) \quad (21.7)$$

where the shorthand notation $A_i^2 = A_i^a A_i^a$ with the upper index is the group index which has been used.

Bali et al. [17] point out another important aspect of this correspondence. The lattice spacing going to zero is equivalent to physical length scales going to infinity in comparison (see Sect. 2.1 of Chap. 20 on lattice gauge theories). Thus an equivalent

limit is for $T \rightarrow \infty$ (so also R). Bali et al. also discuss how to account for the realistic situations of finite T . The energy and action densities (just as in QED) are given by (see Eq. (4) of [20])

$$\epsilon = \frac{E^2 + B^2}{2} \quad \gamma = \frac{E^2 - B^2}{2} \quad (21.8)$$

For static colour distributions, both should be positive, and nearly equal.

21.3 Creutz's Pioneering Numerical Works

Michael Creutz, along with his collaborators, should be credited for opening the grand route to LGT's via Monte Carlo simulations, which were already known for their power and versatility in the study of statistical mechanical problems. By virtue of the remarkable mapping between QFT's on Euclidean lattices and statistical mechanics, it was only to be expected that such Monte Carlo techniques would play a major role in realizing the full potentials of LGT's. Nearly 43 years after Creutz's pioneering study of SU(2) lattice gauge theories in $D = 4$ [15], the wealth of developments brought forth by Monte Carlo techniques has been breathtaking. The reader is urged to refer to the large number of sources cited in our Chap. 20 on lattice gauge theories to get a feel for these exciting developments.

Even prior to his 1980 work, Creutz with his collaborators had already initiated several Monte Carlo investigations into LGT's. These include studies of U(1) gauge theories (compact QED) in $D = 4$, of SU(2) LGT's in $D = 5$, and of *discrete gauge groups* Z_2 , Z_3 and Z_4 in $D = 4$ [40]. It follows from Wilson's work [2] that *all* LGT's confine in the strong-coupling limit. Yet at extreme weak-coupling compact U(1) should get closer and closer to ordinary QED which is certainly not a confining theory. Rather, it is a theory of massless photons. Therefore a second-order or higher order phase transition separating the confined phase from the so-called coulomb phase (also called the spin-wave phase) would be needed for consistency. In the Monte Carlo studies of [14] such a phase transition was indeed seen. The nature of this phase transition is identical to the famous Kosterlitz-Thouless infinite order phase transition in the two-dimensional X-Y model [41]. Likewise, a second-order phase transition separating the strong-coupling confining phase from a weak-coupling non-confining phase was also observed by them in the $D = 5$ SU(2) LGT. They studied the discrete group theories to see evidence for some conjectures of Migdal [3]. The reader may also recall the importance of such discrete LGT's in the light of claims made by Mack and Petkova [42] (see the brief discussion about this in Chap. 20).

On the other hand, such a phase transition appeared to be absent in $D = 4$ SU(2) gauge theory implying that the theory would be confining even at weak coupling. That would have important implications for whether QCD, a SU(3) LGT in $D = 4$, is a confining theory or not. So the prime motivation for Creutz was to address the question whether $D = 4$ confines in the weak coupling or not. Of course, SU(2) is not the gauge group of QCD, but it is a subgroup of the QCD group SU(3). Technically,

at that time, SU(3) Monte Carlo simulations were much harder and Creutz settled for SU(2). That the SU(2) case shares many important features of QCD was shown by Creutz by demonstrating that in SU(2) also, the statistical continuum limit resides at extreme weak coupling.

Creutz's paper is one of the foundational works in LGT and we shall explain its most salient features.

21.3.1 Monte Carlo Simulations

The numerical investigations of Creutz were carried out using the by then well-known methods of Monte Carlo simulations. This is essentially an extension of the Monte Carlo integration method for multi-dimensional integrals. The basic idea is to use the integrand, provided it is positive, as a *probability measure* and sample contributions to the integral by the method of *importance sampling*. In the case of LGT's the dimensionality of the partition function viewed as a multi-dimensional integral is mind-boggling. Though the Monte Carlo techniques have grown in their complexity and sophistication, their essence can still be captured by the elementary treatment we shall be giving. We shall first illustrate the method with the extremely simple case of the one-dimensional Ising model, and then show the generalization for SU(2) LGT's. But first we discuss some important aspects of Monte Carlo method itself.

Though well known, it is worth reiterating the inevitability of resorting to probabilistic methods when dealing with a large number of degrees of freedom. In 1 cc of a gas there are some 10^{23} atoms, each with six continuous degrees of freedom in $d = 3$ (three positions and three momenta for the simplest of atoms representable as point particles)! Let us consider instead a three-dimensional Ising model whose degrees of freedom take only two values $\pm$. Taking, for example, 10 of these in each direction one would have 10^3 d.o.f, staggeringly small compared to 10^{23} ! Nevertheless, with each of them taking two possibilities, the total number of configurations of even this simplest type of system is $2^{1000} \simeq 10^{300}$! Creutz dramatizes the situation by pointing out that even with a mythical hypercomputer with the ability to handle one configuration every 10^{-23} s (no such computer exists even to this day), working for the age of the universe, about 10^{18} s, would only have taken care of 10^{41} configurations, less than miniscule fraction of the total! So, deterministic approaches are doomed from the start.

Let C denote a configuration of some system (in our Ising example above, this would be the 1000 values of Ising spins). Let us consider a probabilistic rule, also called a stochastic process, W , that assigns a probability for every new configuration C' . Of all the possibilities, the so-called *Markov Process* is the simplest to deal with. Colloquially, a Markov process is one where the future is independent of the detailed past, depending only on the present. So the probability $W(C')$ for new configurations does not depend on the *history* of past configurations, but only on the present configuration C . Then it makes sense to denote the stochastic transition from C to C' by $W(C \rightarrow C')$.

It should be appreciated that even changing the value of one spin changes the configuration C . So, a simple example of a Markovian W is to use a random number generator based on the current values of spin to assign probabilities for the two new values of the spin. We shall show how this is done explicitly shortly. Since, W is a probability distribution it must satisfy the obvious requirements of probabilities and their conservation:

$$W(C \rightarrow C') \geq 0 \quad \sum_{C'} W(C \rightarrow C') = 1 \quad (21.9)$$

$W(C \rightarrow C')$ is an example of what are called *Stochastic Matrices*, whose rows and columns are labelled by the initial and final configurations. At least when C is a countable set, one can think of the eigenstates of this matrix, however large. One can extend the discussion also to cases where the degrees of freedom are continuous, though it is technically much much harder. Actually we have already done so in our example of the $D = 1$ transfer matrix discussed in Sect. 2.2 of Chap. 20.

A particularly special eigenstate is designated as the *equilibrium* state $P_{eq}(C)$. The natural question that arises is whether the existence of such an equilibrium state is always guaranteed, and if so, whether it is unique. This is the heart and soul of statistical mechanics! When the positive semi-definiteness of $W(C \rightarrow C')$ is relaxed to strictly positive, i.e. $W(C \rightarrow C') > 0$, a remarkable theorem called the Frobenius-Perron theorem (the reader is referred to the Wikipedia article *Perron-Frobenius Theorem* for an excellent treatment along with important references) in fact guarantees the existence of an unique eigenstate with the largest eigenvalue. This largest eigenvalue eigenstate is the equilibrium state with probability distribution $P_{eq}(C)$. The Markov process W is called an *update* in Monte Carlo parlance and it must leave the equilibrium state unchanged, i.e.

$$P_{eq}(C') = \sum_C W(C \rightarrow C') P_{eq}(C) \quad (21.10)$$

The largest eigenvalue, consequently, is 1. In fact when C, C' are configurations at consecutive (discrete) times, W is essentially the transfer matrix. Before proceeding further, it is important to clarify the meaning of the restriction to strictly positive W , instead of positive semi-definiteness. That means, there is eventual access to every configuration.

We are now in a position to establish an important property about approach to equilibrium. For this we define the concept of a *distance* between two *ensembles* of configurations as

$$||E - E'|| \equiv \sum_C |P(C) - P(C')| \quad (21.11)$$

If E' is obtained from E as a result of the Markov Process W (also called update, algorithm, etc.), it follows that $P'(C) = \sum_{C'} W(C' \rightarrow C) P(C')$, and consequently

$$\begin{aligned}
||E' - E_{eq}|| &= \sum_C |P'(C) - P_{eq}(C)| \\
&= \sum_C \left| \sum_{C'} W(C' \rightarrow C) (P(C') - P_{eq}(C')) \right| \\
&= \sum_C W(C' \rightarrow C) |P(C') - P_{eq}(C')| \\
&\leq \sum_{C'} |P(C') - P_{eq}(C')| \leq ||E - E_{eq}|| \quad (21.12)
\end{aligned}$$

In the second step use was made of $P_{eq}(C) = \sum_{C'} W(C' \rightarrow C) P_{eq}(C')$, and the third step follows from the positive semi-definiteness of W . This means that with each update, the distance to the equilibrium state either decreases or stays the same. Unfortunately, this is not enough to guarantee that the updates eventually take the system to equilibrium. For that, one really needs the Frobenius-Perron theorem (and the strict positivity of W).

These comforting results notwithstanding, the issue of thermalization, i.e. of the system eventually reaching equilibrium, and the equally important issue of how fast thermal equilibrium is reached, remains both conceptually and technically formidable. Alarm bells in this regard have been the *Poincare Recurrence Theorem*, one of the few exact results in Hamiltonian dynamics that follows with nothing more than the compactness of phase space, and the Liouville Theorem [43] on the one hand, and the celebrated *Fermi-Pasta-Ulam* [44,45] work which at sight seemed to invalidate much of the folklore about thermal equilibrium.

A sufficient, but not necessary, condition to ensure that P_{eq} is an eigenstate of W with the largest eigenvalue one is:

$$P_{eq}(C) W(C \rightarrow C') = P_{eq}(C') W(C' \rightarrow C) \quad (21.13)$$

This is called the *Detailed Balance Condition*. Summing both sides over C' and using the second of Eq. (21.9) gives back the eigenvalue condition of Eq. (21.10) for P_{eq} .

Two of the most popular algorithms employed in Monte Carlo studies are: (a) Heat Bath—In this case, one chooses $W(C \rightarrow C')$ to be *independent* of C , i.e. $W(C \rightarrow C') = W(C')$. The detailed balance is trivially satisfied if $W(C)$ is chosen to be proportional to the equilibrium distribution $P_{eq}(C)$. Therefore, in the heat bath algorithm, the final configuration C' is chosen according to $P_{eq}(C')$, irrespective of the current configuration. It is obviously a Markov process.

Often the heat bath algorithm cannot be implemented because $P_{eq}(C)$ is not known in analytical form. For $SU(2)$ it can be implemented, as shown by Creutz [15], and to be discussed shortly. For $SU(3)$ it cannot be implemented very easily. The other popular update algorithm is the so-called *Metropolis Algorithm*. If $S(C)$ is the action for the configuration C , the Metropolis algorithm consists of $W(C \rightarrow C') = e^{S(C') - S(C)}$ if $S(C') < S(C)$; otherwise $W(C \rightarrow C') = 1$. Recall that $P_{eq}(C) \propto e^{\beta S(C)}$. Therefore the Metropolis definitely selects favourable configurations, but

does not altogether reject unfavourable ones. Repeated applications of Metropolis tend to equal Heat Bath updates. Metropolis algorithm is generally time-consuming.

Now the Monte Carlo simulation consists in starting the system in some chosen initial configuration, and sweep the lattice updating each degree of freedom by either heat bath (where possible) or metropolis, and continue till some observable chosen for monitoring equilibrium stabilizes. The initial configurations are usually of two types: (i) cold start—where all the variables are fixed at some value. For example, in the Ising model example, all spins can be chosen to be of one sign. (ii) hot start—in this case, each variable is chosen randomly. The convergence of the system starting from these two diametrically opposite configurations is also used to monitor whether equilibrium has been reached. At that point various observables of interest, including correlation functions, are measured. There are issues of both statistical and systematic errors. We shall not go into them. The interested reader should consult the resources cited.

Heat Bath: We shall illustrate how the heat bath algorithm works for two cases (a) $D = 1$ Ising model and (ii) $SU(2)$ in arbitrary D , as developed by Creutz in [15]. Let us begin with the Ising chain, whose Hamiltonian is given by

$$H = -J \sum_i \sigma_i \sigma_{i+1} \quad \sigma_i = \pm 1 \quad (21.14)$$

This is the classical Ising chain where the variables σ_i are just numbers taking two values ± 1 . Now the idea is to visit each site and update the corresponding spin by the heat bath algorithm, while keeping all other spins fixed. Let us, for example, consider the spin σ_i . Only two terms in the Hamiltonian involve σ_i . They are: $-J(\sigma_{i-1}\sigma_i + \sigma_i\sigma_{i+1})$. Thus the probability distribution for σ_i keeping all other spins fixed is

$$P(\sigma_i) \propto e^{\beta(\sigma_{i-1} + \sigma_{i+1})\sigma_i} \quad \beta = \frac{J}{kT} \quad (21.15)$$

Hence the probabilities for the two possible values of $\sigma_i = \pm 1$ are (with $\Sigma = \sigma_{i-1} + \sigma_{i+1}$):

$$P(\sigma_i = 1) = \frac{e^{\beta \Sigma}}{2 \cosh \beta \Sigma} \quad P(\sigma_i = -1) = \frac{e^{-\beta \Sigma}}{2 \cosh \beta \Sigma} \quad (21.16)$$

In fact Σ acts like an external field. Now the heat bath implementation consists of choosing the values of σ_i according to this distribution. To see how that can be done, let us generate a random number in the range (0,1) uniformly. If the random number happens to lie in the range (0, $P(+)$), we assign the spin value +1 to σ_i , else, the value -1.

Now let us consider the more involved $SU(2)$ LGT in arbitrary dimensions on a hypercubic lattice with the plaquette action described in Chap. 20. Now the Monte Carlo simulations involve essentially the same steps as in the much simpler Ising chain case; sweep the lattice updating the links one by one while keeping all other

links fixed. In the D -dimensional case, there are $D - 1$ planes containing a given direction. Thus $2(D - 1)$ plaquettes will contain any given link. Let us illustrate by considering the link to be updated to be in the 1-direction at some site and denote it as U_1 , the part of the action containing U_1 is of the form

$$\beta \text{Tr } U_1 (S_{1,2} + S_{1,-2} + \dots S_{1,D} + S_{1,-D}) \quad (21.17)$$

Here we have denoted the product of the three links which, along with U_1 , form a plaquette in the $(1, i)$ plane in the clockwise direction as the *staple* $S_{1,i}$ (see Fig. 1 of Chap. 20 for their definition); likewise, the staple in the $(1, i)$ plane in the anti-clockwise direction as $S_{1,-i}$. We are just following the treatment of Creutz, but generalized to arbitrary D whereas he specifically dealt with $D = 4$. It is to be noted that each staple in the above sum, being a product of three $SU(2)$ matrices, is also an $SU(2)$ matrix. As he notes, a special property of $SU(2)$ matrices (not true for $SU(3)$) is that any sum of them is *proportional* to another $SU(2)$ matrix! Thus, the above equation can be rewritten as $\beta k \text{Tr } U_1 \tilde{U}$. The (continuous) probability distribution for U_1 is, therefore,

$$dP(U_1) = c dU_1 e^{\beta k \text{Tr } U_1 \tilde{U}} \quad (21.18)$$

with dU_1 the *Haar Measure* already discussed in Chap. 20. The trick proposed by Creutz is to consider the probability distribution for $V_1 = U_1 \tilde{U}^{-1}$ instead of one for U_1 itself. The invariance property of the Haar measure (see Chap. 20) gives $dU_1 = d(V_1 \tilde{U}) = dV_1$. Therefore the probability distribution for the $SU(2)$ matrix V_1 is:

$$dP(V_1) = c dV_1 e^{\beta k \text{Tr } V_1} \quad (21.19)$$

Using the $SU(2)$ parametrization $V_1 = a_0 + \mathbf{a} \cdot \boldsymbol{\sigma}$ with $a_0^2 + a^2 = 1$, the probability distribution for V_1 can finally be evaluated to be

$$dp(V_1) = e^{2\beta k a_0} \sqrt{1 - a_0^2} da_0 d\Omega \quad (21.20)$$

Now the heat bath implementation is to choose the direction of $\mathbf{a}$ uniformly over the solid angle Ω , and a_0 according to the distribution above. That gives the $SU(2)$ element V_1 ; the updated link is obtained by $V_1 \tilde{U}$. This heat bath algorithm works only for $SU(2)$.

21.3.2 The Statistical Continuum Limit

As discussed in Chap. 20, in LGT's the lattice and the lattice spacing only serve to provide a regularization, albeit a non-perturbative one, of QFT's. The lattice spacing plays the role of a cut-off and just as in the perturbative regularizations of QFT's, the cut-off must eventually be removed. In the perturbative case, the cut-off can be removed only *after* suitable renormalizations. In Chap. 20 on LGT's, we discussed

how the removal of cut-off in LGT's, i.e. of the lattice spacing going to zero, has to be achieved through the statistical continuum limit procedure. This was also tantamount to determining how the bare coupling constant has to be *tuned* with lattice spacing. In [15] Creutz shows how all these are to be done. The particular renormalization scheme that he adopts, called *a renormalization scheme based on confinement*, keeps the string tension (the coefficient of the linear term in the $Q\bar{Q}$ -potential) fixed. We shall describe it more fully. The same renormalization scheme was also used by Kogut et al. [46], but they used the string tension as obtained via strong-coupling expansions, whereas Creutz bases them on string tension measured by his Monte Carlo simulations.

Some general remarks about renormalization schemes on lattices are in order. It typically involves tuning the bare coupling constant keeping some physical quantity fixed. It is also tied up with the issue of whether there are more than one family of correlation lengths that are proportional to each other. In $D = 3$ compact QED, a phase transition separates the confining phase from the coulomb phase. The lattice string tension $\sigma^L(\beta)$ and the lattice mass gap $m^L(\beta)$ have different β -dependences. Thus in principle two physically distinct renormalization schemes are possible; one in which the string tension is held fixed and another in which the mass gap is held fixed.

Creutz also derives the bare coupling constant tuning on the lattice by using the beta function from perturbative QCD expressed in terms of the renormalized coupling. If a phase transition does not separate the strong- and weak-coupling regimes, the tuning of the bare coupling with lattice spacing should remain the same. The conclusion of [46], based on the renormalization scheme keeping the string tension fixed, was that it was indeed the same. Creutz also claims the same based on his Monte Carlo studies.

Let us first take a look at the issue in the perturbative weak renormalized coupling regime. The renormalized coupling constant $g_R(g_0, \mu, a)$ is a function of the bare coupling g_0 , the renormalization scale μ , and the cut-off which in this case is the lattice spacing. The perturbative β -function (Creutz uses the symbol $\gamma(e_R)$ presumably to avoid being confused with β on the lattice, but that still leads to a confusion with γ the anomalous dimension!), to the leading order, is given by

$$\gamma(g_R) \equiv \mu \frac{\partial}{\partial \mu} g_R(g_0(a), \mu, a) \approx -\frac{11}{24\pi^2} g_R^3 + \dots \quad (21.21)$$

The negative sign being reflective of the asymptotic freedom of the theory. Now, g_R being physical is to be held fixed as g_0 is varied wrt a . This defines the lattice renormalization scheme. Consequently,

$$0 = a \frac{d}{da} g_R(g_0(a), \mu, a) = a \frac{dg_0}{da} \frac{\partial e_R}{\partial a} \quad (21.22)$$

Purely on dimensional grounds, $g_R(g_0, \mu, a)$ can only depend on the combination μa from which it follows that

$$a \frac{\partial g_R}{\partial a} = \mu \frac{\partial g_R}{\partial \mu} = \gamma(g_R) \quad (21.23)$$

Creutz combines these relations to arrive at

$$a \frac{d g_0(a)}{d a} = \frac{11 g_0^3}{24\pi^2} + \dots \quad (21.24)$$

where the dots refer to terms higher order in g_0 . This firstly indicates that the bare coupling $g_0(a)$ must decrease as the lattice spacing is reduced. For eventually small g_0 , one integrates this, exactly as was done in perturbative QCD to

$$g_0^2 \approx \frac{12\pi^2}{11 \ln(\frac{\hat{a}}{a})} \quad (21.25)$$

where $\hat{a}$ is an integration constant completely analogous to our discussion of the Λ parameter of QCD in Chap. 19. This can also be rewritten as

$$a^2 \approx \hat{a}^2 e^{-\frac{6\pi^2 \beta}{11}} \quad \beta = \frac{4}{g_0^2} \quad (21.26)$$

This is the lattice analog of asymptotic freedom.

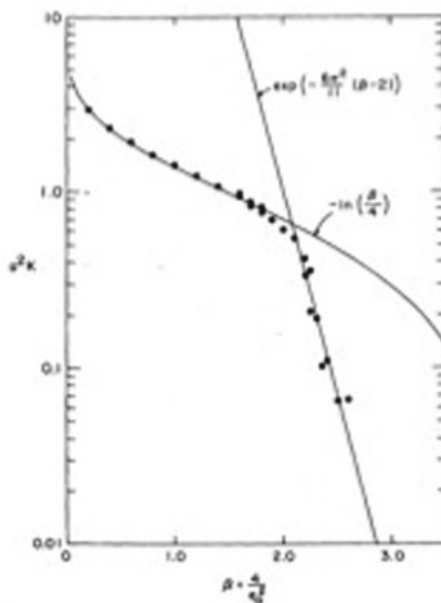
Now we discuss what tuning of g_0 with lattice spacing follows from the non-perturbative renormalization based on confinement [15,46], i.e. where the physical string tension σ_{phys} is held fixed. It's relation to the lattice string tension, based on dimensional analysis, is $\sigma^L(\beta) = \sigma_{phys} a(\beta)^2$. Thus measuring $\sigma^L(\beta)$ through Monte Carlo simulations is tantamount to determining $a(\beta)$. That is

$$a_{nonpert}(\beta) = \sqrt{\frac{\sigma^L(\beta)}{\sigma_{phys}}} \quad (21.27)$$

If no phase transition (second or higher) separates the asymptotic freedom and confinement regimes, $a(\beta)$ as determined by the perturbative analysis and the one determined by fixing the physical string tension must approach each other as one moves towards smaller lattice spacings, equivalently, larger β . The triumph of Creutz's Monte Carlo investigations of $D = 4$ SU(2) LGT is that after about $\beta \approx 2.3$, the non-perturbative and perturbative $a(\beta)$ agreed very well with each other. We have reproduced Fig. 6 of Creutz's paper [15] in Fig. 21.2, in support of this. Needless to say, the numerical resources available to Creutz were miniscule compared to current levels and the errors were much larger. He worked with a 10^4 lattice at the largest $\beta \approx 3.0$. The largest Wilson loops he could measure reliably were around 5×5 .

Nevertheless, the significance of Creutz's work lies on the one hand, in showing a coexistence of asymptotic freedom and confinement in $D = 4$ SU(2) LGT, and on

Fig. 21.2 Creutz's results for string tension of SU(2) gauge theory in four dimensions. Reprinted from [15] ©1980 American Physical Society. Reproduced with permissions. All rights reserved



the other hand, in an approximate determination of the physical string tension, i.e. the coefficient of the linearly rising part of the static quark-antiquark potential. As a result, his work also vindicated the dual superconductor mechanism. As we shall see in the rest of the book, this, despite its enormous importance towards a demonstration of permanent or absolute confinement in non-abelian gauge theories, is only the tip of the iceberg.

21.4 Work of Ambjorn, Olesen and Peterson

Soon after Creutz's work, many important developments took place that impacted the study of flux tubes and confinement on the lattice. On the theoretical side, Nambu had already suggested important parallels between expectation values of large Wilson loops and what may be called *Effective Boson String Theories* [47]. We shall have a lot to say about this connection in the next chapter. Nambu's ideas led Lüscher et al. [48] to study such effective string descriptions further. They found that the linearly rising part of the static potential receives a sub-dominant correction of the form $\frac{a}{R}$ (R is the distance between the quark-antiquark pair) which superficially looks like a coulombic term, but is very different in nature. In subsequent important developments [49,50], Lüscher and collaborators showed that this correction has the *universal* behaviour of the type $-\frac{\pi(D-2)}{24R}$. In string theories, such an universality is easy to understand because of the lack of any free parameters other than the tension. But as effective descriptions of the static potential in LGT's, this is a strikingly surprising result as it only depends on D and not on any aspects of the gauge theory like the nature of the

gauge group, the representation to which the quarks and antiquarks belong, etc. It is in this sense that despite its superficial similarity to coulomb interactions (in $D = 4$), it is fundamentally different. We will have lot more to say on all these aspects in Chap. 22.

In an important sequel to Creutz's 1980 work, Ambjorn et al. [16] undertook a study of Wilson loops of SU(2) lattice gauge theories in $D = 3$ through Monte Carlo simulations. Their choice of $D = 3$ was motivated by several considerations: (i) the ability to go to larger lattices, they could investigate on a $16^2 \times 32$ lattice, (ii) consider Wilson loops with larger temporal extents, they could go up to 6×10 ($R \times T$) loops and finally (iii) to go to much larger values of $\beta = 6.5$. According to their studies of Creutz ratios, scaling behaviour could be seen to set in around $\beta = 4.5 - 5.0$. To speed up the simulations they used the largest subgroup of SU(2) with 120 elements (the Icosahedral group), but the disadvantage of this is that beyond $\beta = 6.5$ the group dynamics freezes. Yet another reason they cite for $D = 3$ is that the coulombic interaction not being of the $\frac{1}{R}$ form can be distinguished more easily from the universal Lüscher term.

Ambjorn et al. compare their results to the theoretical expectations from a string model that was termed the Nielsen-Susskind type in our Chap. 17. The reader is referred to their Eq. (4). In order that the dominant linear term in the potential does not completely obscure the sub-dominant terms, they consider, instead of $V(R)$ (it is worth emphasizing that their theoretical results for $\langle W \rangle$, for finite R, T , have terms like $\ln R\mu$), the combination $Q(R) = V(R) + V(R-2) - 2V(R-1)$ and plot $\frac{R(R-1)(R-2)}{2} Q(R)$ which, for very large R , approach the scaled second derivative $-\frac{R^3}{2} V''(R)$. This is also what Lüscher and Weisz followed (to be discussed later in the chapter), and also us (Pushan Majumdar and N.D. Hari Dass). This way, not only the constant and linear terms of the potential do not show up, also the coulombic term in $D = 3$ which is $\ln(R)$ would have a linearly growing contribution! Furthermore, for very large R and large T , it should approach the negative of the coefficient of the linear term. In their Fig. 21.2, they have shown data for $R = 4$ for various values of T . Even at $T = 12$, it approaches 0.10, whereas the Lüscher term value would have been $\frac{\pi}{24} \approx 0.131$. Thus the work of Ambjorn et al. has taken the static potential one important step in the right direction, namely, the linearly rising term is accompanied by a sub-dominant term, most likely of the Lüscher term type. The value $R = 4$ is not large enough to say more.

21.5 Flux Profile Studies

Nearly a decade after the work of Ambjorn et al., two groups simultaneously and independently carried out extensive studies of the flux tube profiles. One was Bali et al. [17,18], and the other, Haymaker et al. [19,20]. The earliest investigations of flux tube profiles, and the use of plaquette-Wilson loop correlation functions towards this end can be found in [39]. Bali and coworkers studied both the flux tube profiles in terms of energy density and action density, as well as the static potential while Haymaker et al. only studied the flux tube profiles. Both the groups

studied SU(2) LGT in $D = 4$. Haymaker and collaborators used a $17^3 \times 20$ lattice at $\beta = 2.3, 2.4, 2.5$. They claimed to see flux tube formations starting at separations of around 1.0 fm (a physical scale like fermi is introduced through what is called a Sommer scale).

Bali et al. worked on lattices of size 16^4 , 32^4 and $48^3 \times 64$, the last two much larger than what was used by Haymaker et al. The β -values of Bali et al. were also higher at 2.5, 2.635 and 2.74. According to them string formation could be established over distances as large as 2 fm. In addition to the flux profiles, they also obtained the static potential to unprecedented accuracies. They used two different parametrizations for the static potential given by

$$V(R) = V_0 + K R - \frac{e}{R} \quad V(\mathbf{R}) = V_0 + K R - \frac{e}{R} + f\left(\frac{1}{R} - 4\pi G_L(\mathbf{R})\right) \quad (21.28)$$

where G_L is the lattice gluon propagator computed on an infinite lattice. Thus the f term is sub-dominant even to the $\frac{e}{R}$ term. They use several different fitting methods. For their three parameter fits (V_0, K, e) at the largest $\beta = 2.74$ on 32^4 lattice, the best fit for e is stable at approximately 0.2176 as compared to the value of 0.262 expected for $D = 4$. Their four parameter fit (V_0, K, e, f) at $\beta = 2.635$ on their biggest lattice of $48^3 \times 64$ gives values for e around 0.262! The coefficient f is comparable to e . Therefore their analysis already hints at a significant correction to even the Lüscher term.

Both the works have voluminous data and many sophisticated techniques for noise reduction. They both make use of the so-called Michel sum rules. We shall not go into those details here. The flux tube profiles obtained by Bali et al. are much sharper. We conclude by displaying in Fig. 21.3 their data for the static potential, scaled by $\sqrt{K}$, compared to the expectation $V(R) = K R - \frac{\pi}{12R}$ for four choices of lattices; that within error bars, the four data fall on each other is a strong indication that continuum limit has been reached well. We also display in Fig. 21.4 their action density profile for $R = 14$ (corresponding to a separation of 1.2 fm) at $\beta = 2.51155$. A noteworthy

Fig. 21.3 The static potential of [17] ©1995 The American Physical Society. Reproduced with permissions. All rights reserved

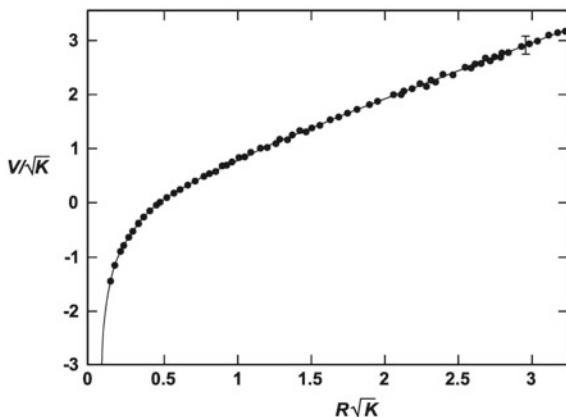
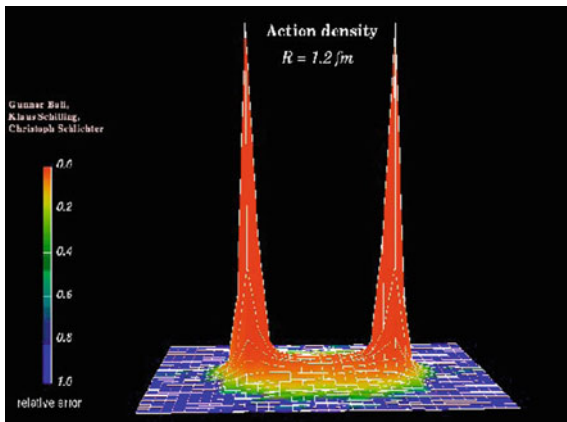


Fig. 21.4 The action density profile of flux tubes. Reproduced with the kind permission of Prof. G. Bali. All rights reserved



feature of the profiles obtained by both the groups is that the flux tubes are *thick*, a feature that is very important to understand.

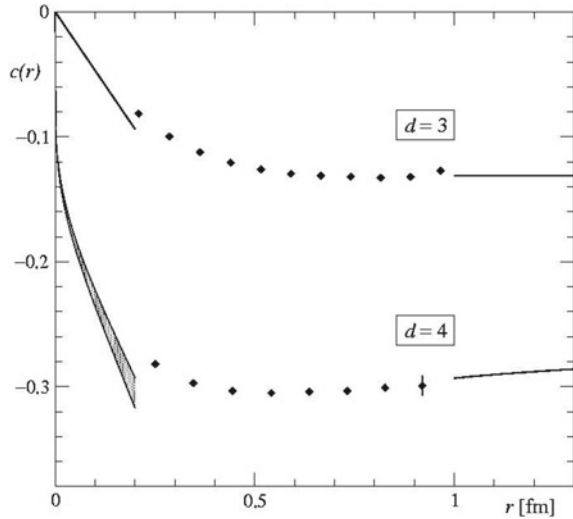
In accordance with the adage “a picture is worth a million words”, this figure above captures all the essentials of our book! It not only vindicates the idea of the dual superconductor mechanism as what underlies *Quark Confinement*, it also captures the *Coulomb*-like behaviour at short distances, characteristic of the asymptotic freedom regime of the gauge theories. A different aspect of this figure will later be provided by the high accuracy results on the static potential.

21.6 Simulations of Lüscher and Weisz

A major development in the study of the static potential was ushered in by Lüscher and Weisz. Instead of Wilson loops they opted for the much harder Polyakov Loop Correlators for the static potential determination. As explained in the early part of this chapter, it is equivalent to studying Wilson loops with the largest possible temporal extent. Also, while this provides the best possible projection onto the ground state (in the presence of the static $Q\bar{Q}$ pair), the signal gets exponentially (in T) suppressed. As a result, it becomes a challenge to control the noise (read statistical errors).

This was made possible by an amazing algorithm invented by them called the *multilevel* [21]. This algorithm exploits locality of actions (the Wilson plaquette action being one such) and the so-called *sublattice updates* to achieve *exponential* reduction in statistical errors. In their paper, they specifically show how this can be achieved for Polyakov Loop correlators of $D = 4$ SU(3) LGT's and that accuracies can be improved by several orders of magnitude when the areas bounded by the loops become very large. We shall not go further into the details but strongly recommend the reader to their paper. In fact, the multilevel algorithm should find applications in Monte Carlo simulations in other areas, particularly when ultra-non-local observables need to be investigated numerically.

Fig. 21.5 Results of Lüscher and Weisz [22]
 ©SISSA/ISAS 2002.
 Reproduced with
 permissions. All rights
 reserved



In their pioneering studies based on Polyakov Loop correlators as well as the powerful multilevel algorithm titled *Quark Confinement and the Bosonic String* [22], Lüscher and Weisz studied SU(3) LGT's in both $D = 3, 4$. The β values for their studies were 11.0 on a $32^2 \times 24$ lattice, 15.0 on a $48^2 \times 32$ lattice and 20.0 on a $60^2 \times 36$ lattice for $D = 3$ and 5.7 on a $24^3 \times 18$ lattice, 5.85 on a $36^3 \times 24$ lattice and 6.0 on a $48^3 \times 30$ lattice for $D = 4$. Their results for their finest lattices at $\beta = 6.0$ for $D = 4$ and $\beta = 20.0$ for $D = 3$ are shown in Fig. 21.5. They determined the Lüscher term to nearly 15% accuracy. We shall address their proposals to account for this discrepancy in the next Chap. 22. It is important to keep in mind that the largest $Q\bar{Q}$ -separations they could realize were about 0.82 fm for $D = 4$, and less than 1 fm for $D = 3$.

In the figures, the solid curves to the right represent the potential including the Lüscher term and a boundary contribution with a best fit. What is plotted on the vertical axis is $c(r) = -\frac{r^3}{2} V''(r)$, which was also used by Ambjorn et al. The left side of the figure shows the result of perturbative QCD to two loops in $D = 4$, and to one loop in $D = 3$ (as explained in Chap. 19, in three dimensions the perturbative calculations are *infrared* divergent beyond one loop). The analytical expressions used by Lüscher and Weisz are given below:

$$c(r) = -\frac{1}{3\pi} g^2 r + \dots \quad (21.29)$$

for $D = 3$ SU(3) case, and

$$c(r) = -\frac{4}{3} \alpha\left(\frac{1}{r}\right) \quad \alpha(\mu) = \frac{\bar{g}(\mu)^2}{4\pi} \quad (21.30)$$

valid for two loops. The author is not able to reproduce this expression for $D = 4$; it looks very different from the explicit derivations given in Sect. 19.7 of Chap. 19. The

$D = 3$ expression used by Lüscher and Weisz follows from the work of Schröder [51] who showed that beyond 1-loop, the static potential develops infrared singularities. His expression for the two-loop static potential is

$$V(r) = \frac{g^2 C_F}{2\pi} \ln g^2 r + \frac{7 g^4 C_F C_A}{64\pi} \cdot r \quad (21.31)$$

where the first term is the one-loop result and the next term the two-loop correction. For $SU(N)$ theories, $C_F = \frac{N^2 - N}{2N}$ and $C_A = N$ (see Chap. 19). Hence in $D = 3$ the two-loop correction would give $c(r) = 0$ (for all $SU(N)$ cases), while the one-loop correction above gives $c(r) = -\frac{C_F}{4\pi} g^2 r$. For $SU(3)$ for which $C_F = \frac{4}{3}$, this yields $c(r) = -\frac{g^2 r}{3\pi}$, the expression used by Lüscher and Weisz.

The most striking conclusion of Lüscher and Weisz's work is not only that the asymptotic freedom and confinement coexist in non-abelian gauge theories, but also that the transition takes place rather quickly. We will have more to add to this in the next section.

21.7 Simulations of Hari Dass and Pushan Majumdar

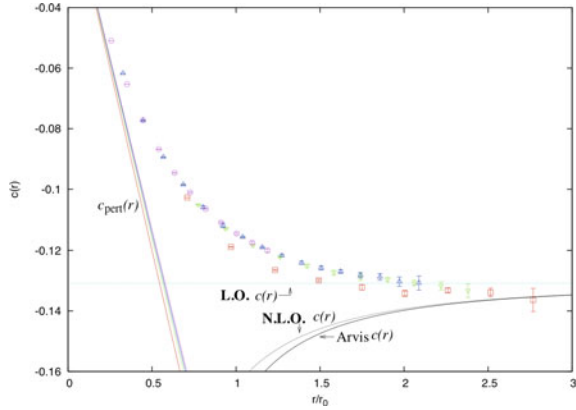
In 2004 the author had built a 144 dual CPU node teraflop Linux cluster (details can be found in [52]). With Pushan Majumdar, who unfortunately passed away at a very young age 2 years ago [33], I decided to extend Lüscher and Weisz's studies of flux tubes for both $SU(2)$ and $SU(3)$ in $D = 3, 4$. The design features of this supercomputer KABRU were maximally utilized to study fairly big lattices as well as much larger $Q\bar{Q}$ -separations. We also based our investigations on Polyakov Loop correlators and the multilevel algorithm. Increasing the separations by even one lattice spacing adds tremendous numerical burdens!

The details of our investigations can be found in [23–26]. In $D = 4$ we studied the $SU(3)$ case on a 32^4 lattice at $\beta = 5.7$ [23, 24]. This is, of course, coarser than the finest lattice used by Lüscher and Weisz. Our choice was dictated by the desire to go to larger $Q\bar{Q}$ separations and the hope that at such large separations the relative coarseness of the lattice would not matter. Thus we could go to separations as large as 1.2 fm in comparison to the maximum of 0.82 fm by Lüscher and Weisz. We shall soon see the importance of going beyond 1 fm. We had however started runs at $\beta = 6.0$ on a $48^3 \times 32$ lattice which could not be completed.

In $D = 3$ we studied the $SU(2)$ case at beta value of 5.0 on $36^3, 40^3, 48^3$ lattices, at 7.5 on $48^3, 64^3$ lattices, at 10.0 on $48^3, 84^3$ lattices and finally at 12.5 on $48^3, 72^3$ lattices [25, 26]. In this case we could realize separations as large as 1.37f as compared to the Lüscher-Weisz values of around 1 fm. Our results for the scaled second derivatives $c(r)$ are shown in the next two figures.

Let us first discuss the $D = 3$ $SU(2)$ case, shown in Fig. 21.6. The four solid straight lines are the results from two-loop perturbation theory, as given in [51]. Now $C_F = \frac{3}{4}$, and the one-loop term yields the sole contribution $c(r) = -\frac{3}{16\pi} g^2 r$. The two-loop result of [51], being a linear term in $V(r)$, does not contribute to $c(r)$. One

Fig. 21.6 Results for $D = 3$
 SU(2). Reprinted from
 [26]©2007 Elsevier.
 Reproduced under
 CC-BY-4.0 License



sees that even at rather short distances of the order of 0.1 fm, the lattice results begin to deviate significantly from the results of [51]. The high accuracy data of [25,26] was used to fit the so-called infrared counterterms to yield an IR-improved static potential of the form (based on dimensional grounds; in $D = 3$ the gauge coupling has dimensions of $[M]^{\frac{1}{2}}$):

$$V_{pert}^{D=3} = \frac{g^2 C_F}{2\pi} \ln g^2 r + \frac{7C_F C_A g^4}{64\pi} r + A g^4 r \ln g^2 r + B g^6 r^2 + \dots \quad (21.32)$$

A and B were fitted to the values $A = 0.013162(3)$, $B = 0.001089(1)$. The IR-improved two-loop results extend the validity of the perturbation theory from 0.1 to 0.25 fm.

At larger distances, the data displays good convergence towards what can be called the *truncated Arvis Potential*. Recall from our Chap. 17 the Arvis quantization of strings with fixed ends. The potential (ground state energy) so obtained was

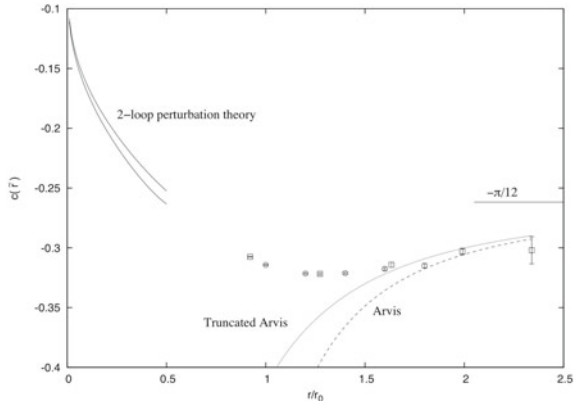
$$V_{arvis}(R) = \sqrt{\sigma^2 R^2 - \frac{(D-2)\pi\sigma}{12}} \quad (21.33)$$

For large R , i.e. $R \gg R_c = \sqrt{\frac{(D-2)\pi}{12\sigma}}$, and keeping the first three terms one gets the truncated or NLO Arvis potential

$$V_{trunc}^{arvis}(R) = \sigma R - \frac{(D-2)\pi}{24 R} + \frac{(D-2)^2\pi^2}{8 \cdot 144 \cdot \sigma R^3} + \dots \quad (21.34)$$

The curves on the right are the truncated Arvis and the full Arvis potentials. Thus here too a coexistence of asymptotic freedom and confinement is seen with a very smooth interpolation between them. Except for the crudest case of $\beta = 5$, the other data points fall on the same curve, pointing to a good convergence to the continuum limit. What is plotted is the dimensionless ratio $\frac{R}{r_0}$ where $r_0 = 0.5$ fm is the so-called Sommer scale.

Fig. 21.7 Results for $D = 4$ SU(3). Reprinted from [24] ©SISSA 2006. Reproduced with permissions. All rights reserved



Coming to the $D = 4$ SU(3) results, shown in Fig. 21.7, the two solid curves on the right side are two-loop perturbative QCD results (two, because of the uncertainties in the so-called Λ -parameter; a better way of displaying the short-distance results is not to evoke the Λ parameter). As already stressed in Chap. 19, perturbative calculations of the static potential in $D = 4$ suffer from infrared divergences at three loops and beyond [53]. Again, as stressed in Chap. 19, there are several schemes (though all inter-related) for calculating the static potential, the so-called $V, \bar{V}, F$, and FS schemes [54–56]. In Chap. 19, we have given explicit derivation of the static potential $V(r)$ and $c(r)$ in the F -scheme. We reproduce here the result for $D = 4$ $c(r)$ of Eq. (19.63):

$$c(r) = -C_F \{(\alpha_{\overline{MS}}(r) + c_0 \alpha_{\overline{MS}}^2(r)) - \frac{r}{2} \partial_r (\alpha_{\overline{MS}}(r) + c_0 \alpha_{\overline{MS}}(r)^2)\} \quad (21.35)$$

The expressions for c_0 and $\alpha_{\overline{MS}}(r)$ are given in Chap. 19. Unfortunately, our lattice was too crude (the lattice spacing was 0.17 fm) so no measurements of the static potential in this range ($R < 0.25$ fm) could be made. However, beyond 1 fm a clear convergence is seen to both the Arvis and truncated Arvis potentials (at such large R the two do not differ appreciably). The biggest surprise from these simulations was the coefficient of the $\frac{1}{R^3}$ term of the static potential is also *universal* just like the Lüscher term.

However, a big puzzle remains! Expressions like the Lüscher term as well as the Arvis potential, derived from string theories, are only consistent in $D = 26$ as elaborated at length in Chap. 17. So the use of $D = 3, 4$ in such formulae can hardly be justified. A partial resolution is to take the attitude that the string models used are only *effective*. That way the use of the Lüscher term for D other than 26 is justified. But that still leaves unanswered the question as to why the truncated Arvis potential can also be used away from 26 dimensions. These important issues are analysed and resolved in Chap. 22.

References

1. J. Schwinger, Phys. Rev. **125**, 397 (1962); Phys. Rev. **128**, 2425 (1962)
2. K. Wilson, Phys. Rev. D **10**, 2445 (1974)
3. A.A. Migdal, Z. Eksp. Teor. Fiz. **69**, 810 (1975); **69**, 1457 (1975); Sov. Phys. JETP **42**, 413 (1975)
4. G. 't'Hooft, High Energy Physics Editorice Compositori Bologna (1975)
5. S. Mandelstam, Phys. Rep. **C23**, 245 (1976)
6. Y. Nambu, Phys. Rev. D **10**, 4262 (1974)
7. Y. Nambu, Phys. Rep. **C23**, 237 (1976)
8. H.B. Nielsen, P. Olesen, Nuc. Phys. **B61**, 45 (1973)
9. G. Parisi, Phys. Rev. D **10**, 870 (1975)
10. M. Creutz, Phys. Rev. D **10**, 2696 (1974)
11. A. Jevicki, P. Senjanovic, Phys. Rev. D **11**, 860 (1975)
12. J. Greensite, *An Introduction to The Confinement Problem*. Lecture Notes on Physics, vol. 821 (Springer, 2011)
13. G. Ripka, arXiv:0310102(hep-ph)
14. M. Creutz, Phys. Rev. Lett. **43**, 553 (1979)
15. M. Creutz, Phys. Rev. D **21**, 2308 (1980)
16. J. Ambjorn, P. Olesen, C. Peterson, Phys. Lett. **142B**, 410 (1984)
17. G.S. Bali, C. Schlichter, K. Schilling, Phys. Rev. D **51**, 5165 (1995)
18. G.S. Bali, Phys.Rep. 343 (2001)
19. R.W. Haymaker, J. Wosiek, Phys. Rev. D **43**, 2676 (1991)
20. R.W. Haymaker, V. Singh, Y. Peng, J. Wosiek, Phys. Rev. D **53**, 389 (1996)
21. M. Lüscher, P. Weisz, JHEP **09**, 010 (2001)
22. M. Lüscher, P. Weisz, JHEP **0207**, 049 (2002)
23. N.D. Hari Dass, P. Majumdar, PoS(LAT2005) 312
24. N.D. Hari Dass, P. Majumdar, JHEP **0610**, 020 (2006)
25. N.D. Hari Dass, P. Majumdar, PoS Lattice **2007**, 316 (2007)
26. N.D. Hari Dass, P. Majumdar, Phys. Lett. B **658**, 273 (2008)
27. K.J. Juge, J. Kuti, C. Morningstar, Phys. Rev. Lett. **90**, 161601 (2003)
28. J. Kuti, Proc. Lattice 2005, PoS(Lat2005), 1 (2005)
29. M. Casselle, M. Pepe, A. Rago, JHEP **0410**, 605 (2004)
30. B.B. Brandt, P. Majumdar, Phys. Lett. B **682**, 253 (2009)
31. B.B. Brandt, P. Majumdar, JHEP **029**, 040 (2011)
32. B.B. Brandt, Indian. J. Phys. (August 2021) **95**(8), 1613 (2021)
33. N.D. Hari Dass, Indian. J. Phys. (August 2021) **95**(8), 1591 (2021)
34. S.L. Adler, Nuc. Phys. **B217**, 381 (1983)
35. M. Baker, J.S. Ball, F. Zachariasen, Phys. Rev. D **34**, 3894 (1986)
36. M. Baker, J.S. Ball, F. Zachariasen, Phys. Rep. **209**, 73 (1991)
37. M. Baker, J.S. Ball, F. Zachariasen, Phys. Rev. D **41**, 2612 (1990)
38. E. Seiler, *Gauge Theories as a Problem of Constructive Quantum Field Theory and Statistical Mechanics*. Lecture Notes in Physics, vol. 159 (Springer, 1982)
39. M. Fukugita, T. Niuya, Phys. Lett. B **132**, 374 (1983); J.W. Flower, S.W. Otto, Phys. Lett. B **160**, 128(1985); R. Sommer, Nuc. Phys. B **291**, 673 (1987); Nuc. Phys. B **306**, 180 (1988)
40. M. Creutz, L. Jacobs, C. Rebbi, Phys. Rev. Lett. **42**, 1390 (1979); Phys. Rev. D **20**, 1915 (1979)
41. J.M. Kosterlitz, G.J. Thouless, J. Phys. **C6**, 1181 (1973)
42. G. Mack, V.B. Petkova, Ann. Phys. NY **23**, 442 (1979)
43. V.I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, 2013)
44. E. Fermi, J. Pasta, S. Ulam, M. Tsingou, *Studies of Nonlinear Problems I*, Los Alamos Preprint, LA-1940 (1955)
45. J. Ford, The Fermi-Pasta-Ulam problem: paradox turns discovery. Phys. Rep. **213**(5), 271 (1992)
46. J.B. Kogut, R.B. Pearson, J. Shigemitsu, Phys. Rev. Lett. **43**, 484 (1979)
47. Y. Nambu, Phys. Lett. **80B**, 372 (1979)

48. M. Lüscher, K. Symanzik, P. Weisz, Nuc. Phys. **B173**, 365 (1980)
49. M. Lüscher, Nuc. Phys. B 180[FS2] 317 (1981)
50. M. Lüscher, G. Münster, P. Weisz, Nuc. Phys. B 180[FS2], 1 (1981)
51. Y. Schröder, The static potential in QCD, DESY-THESIS-1999-021; Phys. Lett. B **447**, 321 (1999)
52. N.D. Hari Dass, *Lattice Gauge Theory and KABRU - the Teraflop Linux Cluster at IMSc*, 107, Recent Trends in Practice and Theory of Information Technology, Proceedings of NRB Seminar, NPOL Kochi, 10-11 January 2005, ed. by S.N. Maheshwari
53. N. Brambilla, A. Pinoda, J. Soto, A. Vairo, Phys. Rev. D **60**, 09152 (1999)
54. S. Necco, R. Sommer, Phys. Lett. B **523**, 135 (2001)
55. S. Melles, Phys. Rev. D 074019 (2000)
56. A. Laschka, N. Kaiser, W. Weise, *IX International Conference on Quark Confinement and Hadron Spectroscopy(QCHS)* (Madrid, 2010)

Flux Tubes and Effective String Theories (EST)

22

22.1 Introduction

In Chap. 21 we discussed in detail the numerical investigations of flux tubes. The investigations till around 1995 had established rather accurately the Lüscher term in the static $Q\bar{Q}$ -potential:

$$V(R) = \sigma R - \frac{(D-2)\pi}{24 R} + \dots \quad (22.1)$$

The remarkable feature of the Lüscher term is its *universality*. It only depends on the space-time dimension D , and on no other details of the gauge theory like the structure of the gauge group, the representation to which quarks belong etc. The accuracy in the determination of this potential was further increased by the very high accuracy studies of Lüscher and Weisz in 2002 [1] who utilised the path-breaking *Multilevel* algorithm [2], which can deliver exponential reduction in statistical errors. Lüscher and Weisz also analysed these results as well as the spectrum [1, 3]. These studies were extended further by Hari Dass and Majumdar [4–7] who too made use of the multilevel algorithm to push investigations in $D = 3$ SU(2) and $D = 4$ SU(3) lattice gauge theories to larger $Q\bar{Q}$ -separations than before. They found reasonable evidence for universality of even higher-order (in $\frac{1}{R}$) corrections to Eq. (22.1). In particular, they found that the data in both $D = 3$ and $D = 4$ favour the *truncated Arvis Potential*:

$$V_{arvis}^{trunc}(R) = \sigma R - \frac{(D-2)\pi}{24 R} - \frac{(D-2)^2\pi^2}{1152 \sigma R^3} + \dots \quad (22.2)$$

Of course, both the accuracies as well as the largest $Q\bar{Q}$ -separations probed by Hari Dass and Majumdar need to be pushed even further to confirm that this is indeed the behaviour of the static potential.

The coefficient of the term linear in R , σ , is called the string-tension. In the Monte Carlo studies of the flux tube described in the previous Chap. 21, the lattice renormalization scheme adopted was the one in which this physical string tension was held fixed. The difference between physical and lattice quantities is explained in Chap. 20 on Lattice Gauge Theories, as well as in Chap. 21.

A puzzling feature here, and something that bothered this author in the beginning, is that these expressions were working well even for $D = 3, 4$ while their derivations [8], being based on bosonic string theory, should only be valid for $D = 26$. The derivation of the full Arvis potential that yields the above when truncated to order $\frac{1}{R^3}$ as well as the $D = 26$ restrictions are discussed at length in chapter 17. The calculation of the Lüscher term was also based on the concept of effective strings(bosonic), inspired by a parallel found by Nambu between expectation values of large Wilson loops and certain string theoretic constructs [9].

The main objective of this chapter is to discuss such effective string theory descriptions of these results that also explain why it is legitimate to use the above-mentioned results even when $D \neq 26$. Before going into the details, it is important to clarify the concept of an effective string theory. It is more or less in the same spirit as the *Effective Field Theories* that we discussed in Chap. 18, in particular the effective field theories of strong interactions. The spirit there was to construct field theories with some effective low energy degrees of freedom in such a way that the symmetry content of the anticipated microscopic theory (in this case the $SU(2) \times SU(2)$ symmetry of QCD) was preserved in the effective theory, but such requirements as *renormalizability* or equivalently, good high energy behaviours are no longer required. In that spirit, it could be hoped that effective string descriptions exist which retain the symmetry content of string theories, but only restricted to such physical circumstances (to be specific, the large distance behaviour) that issues like restrictions to $D = 26$ are no longer warranted.

Indeed such a clue is provided by the full form of the Arvis potential itself (reader is urged to check details discussed in Chap. 17):

$$V_{arvis}(R) = \sqrt{\sigma^2 R^2 - \frac{(D-2)\pi}{12} \sigma} \quad (22.3)$$

Below a critical $R_c = \sqrt{\frac{(D-2)\pi}{12\sigma}}$, the potential turns complex reflecting the *tachyonic instability* of the bosonic string theory. On the other hand, no such instability can occur in the full theory i.e. QCD which is a perfectly unitary theory. Therefore, any effective string description can only be expected to be a reasonable description only for large R . The other notable feature of both the Lüscher term and the Arvis potential is the dependence on D being only through the combination $D - 2$. In string theory (see Chap. 17 for an extensive discussion of this important point) this is a reflection of only the transverse vibrations being physical. That in turn is a consequence of the world-sheet general coordinate invariance of the Nambu-Goto action. Since the Nambu-Goto action does not involve any world-sheet metric, this is a true *symmetry* unlike general coordinate invariance in General Relativity where there is an intrinsic metric. We have alluded to this important distinction between

symmetries and invariances already in Chap. 18. In the case of the Nambu-Goto action for string theory, this subtle but important distinction is not often appreciated. In the context of effective string theories, it is this symmetry that plays a significant role (as did chiral symmetry, for example, in the effective description of strong interactions).

While we have mostly emphasized the static-potential which corresponds to the ground state of a string with both ends fixed, the Arvis potential has an easy generalization to the excited states. We record that here for the sake of discussions later on:

$$E_{n,arvis}(R) = \sqrt{\{\sigma^2 R^2 + 2\pi\sigma[-\frac{(D-2)}{24} + n]\}} \quad (22.4)$$

Again, for use in later parts of this chapter, we truncate this also to order $\frac{1}{R^3}$ and rearrange the terms to write it as

$$E_{n,arvis}^{trunc}(R) = V_{arvis}^{trunc} + \frac{\pi}{R} n + \frac{\pi^2}{R^3} \cdot \frac{1}{2\sigma} n \left(\frac{(D-2)}{12} - n \right) \quad (22.5)$$

The basic variables of effective string theories are also $X^\mu(\xi^a)$ as in the string theories described in Chap. 17, with ξ^a the world-sheet coordinates, usually denoted by σ, τ , though with our current notation σ for string-tension this is bound to cause some confusion! In what follows, we first discuss effective string theories as discussed by Lüscher and Weisz(LW) in their 2002 work, and then the effective string theories first propounded by Polchinski and Strominger(PS), and further extended by the author, along with his coworkers Peter Matlock and Yashas Bharadwaj. In the parlance of string theory(again see Chap. 17 for details), the Lüscher-Weisz effective theory is an example of the *non-covariant* approach, while the Polchinski-Strominger approach is that of the *covariant* approach. The Lüscher-Weisz choice can also be described as the *static* gauge (also *orthogonal* gauge). But such a nomenclature is tied up with the symmetries of the classical action; in the string theory described in Chap. 17, the classical action was the Nambu-Goto action and the symmetries were the world-sheet general coordinate invariances. It is only such underlying symmetries or invariances that can guarantee the equivalence of such approaches. A case in point is gauge invariance, which is not a symmetry, but nevertheless guarantees the equivalence of descriptions with different choices of gauges. Another important difference is that while LW use the path-integral formulation with a classical action as input, the PS approach is canonical (operator based).

22.2 Lüscher-Weisz Effective String Theories

The effective Bosonic string theory adopted by Lüscher and Weisz is described first in their 2002 paper [1] and elaborated in great detail in their 2004 paper [3]. Following Nambu's ideas, they express the Polyakov loop correlators as a functional integral for an effective string theory governed by an action. They take this functional integral to be over all two-dimensional string world-sheets bounded by rectangular boundaries

composed of the two Polyakov loops. When R and T are very large (the scale is set by the string tension σ), this string functional integral is dominated by the so called *minimal surface*, which is spanned by the two Polyakov loop(line)s as well as *straight lines* along a spatial direction. Physically this amounts to a string configuration that is a rigid rod. The rest of the functional integral is evaluated by standard techniques for integrating over the fluctuations from the minimal surface whose area is just RT .

Lüscher and Weisz characterize the fluctuations by only the $D - 2$ transverse degrees of freedom. This is by no means obvious in a effective string theory. Of course the Lüscher term only depends on $D - 2$ pointing to some justification for this, but it is by no means obvious that higher corrections to the Lüscher term also imply this. In fact, as shown by Dietz and Filk [10], a large class of string actions give rise to the same Lüscher term making it not a particularly sensitive test of effective string theories. This will become clearer as we go along. As seen from our discussions in Chap. 17, the transversality of the physical degrees of freedom is really a consequence of the symmetry of the action and furthermore, that action (the Nambu-Goto action) is proportional to the area of the world sheet. That Lüscher and Weisz state that the minimal surface dominates the functional integral also points to a classical action that is just the area of the world sheet. But they do not explicitly state this. This is important as many of their results using the quantum spectrum can actually be understood directly in terms of the symmetries of the classical action in so far as results to order $\frac{1}{R^3}$ are concerned. More on this later.

22.2.1 Leading Order Analysis

In their 2002 paper itself LW set up the essentials of their effective string theory. They label the world-sheet coordinates by (z_0, z_1) , and their $(D - 2)$ transverse fields $h(z) = (0, 0, h_2(z) \dots h_{D-1}(z))$. The string is taken to be fixed at the two ends $z_1 = 0, R$. This amounts to the transverse fields obeying Dirichlet boundary conditions at the two end-points. They take their leading action to be of the form

$$S_{eff} = \frac{1}{2} \int_0^T \int_0^R dz_0 dz_1 \partial_a h \cdot \partial_a h + \dots \quad (22.6)$$

where $\cdot$ denotes the ‘scalar product’ in transverse space. They say this form is dictated by symmetry considerations but do not specify what those are. The author guesses they have the two-dimensional Lorentz invariance in mind. Accordingly they say the higher order terms are Lorentz-invariant combinations of derivatives of h . In their treatment, h is dimensionless reflecting the fact that the actual fluctuations have been scaled by $\sqrt{\sigma}$. This is also reflected in the dimensionless coefficient $\frac{1}{2}$ for the leading term. Of course, the higher order terms will necessarily have dimensionful coefficients rendering the theory *non-renormalizable*. We shall have more to say on this as we go along.

Lastly, they assume periodic boundary conditions along the z_0 -direction, i.e. $h_i(T, z_1) = h_i(0, z_1)$. The leading order string functional integral can be evaluated exactly and has been known for a long time in string-literature (see their paper

for a list of sources; also Polchinski's vol I on String Theory [11]). This then leads to their exact result at this order

$$\langle P(x)^* P(y) \rangle = e^{-\sigma R T - \mu T} [\det(-\Delta)]^{-\frac{(D-2)}{2}} \quad (22.7)$$

where Δ is the Laplacian on a two-dimensional cylinder of height R and circumference $2\pi T$ with Dirichlet boundary conditions. The quantity μ is a free parameter arising essentially out of renormalization effects. It already made its appearance in the results of Ambjorn et al. [12]. Fortunately, it only contributes an additive constant to the static potential and is hence devoid of any physical significance. As LW point out, such terms arise in the gauge theory itself arising out of the need to renormalize Wilson loop expectation values as well as Polyakov loop correlators. It is however curious that neither the Arvis quantization of fixed-end strings with the Nambu-Goto action, nor the Polyakov-Strominger(PS) effective string description presented in the next section, require such explicit renormalizations.

The result for the determinant is $\det(-\Delta) = \eta(q)^2$ where $\eta(q)$ is the Dedekind- η function, given explicitly by

$$\eta(q) = q^{\frac{1}{24}} \prod_{n=1}^{\infty} (1 - q^n) \quad q = e^{-\pi \frac{T}{R}} \quad (22.8)$$

For large T , one can expand in powers of q which straight away yields the spectral decomposition for the Polyakov loop correlators! The result for the energy levels is

$$E_n(R) = \sigma R + \mu + \frac{\pi}{R} \left\{ -\frac{(D-2)}{24} + n \right\} \quad (22.9)$$

The first few degeneracies are given by $w_0 = 1$, $w_1 = (D-2)$, $w_2 = \frac{1}{2}(D-2)(D+1)$ (this should be compared with results from the oscillator representation of dual resonance models worked out as an example in Chap. 16). Comparison with Eq. (22.1) shows complete agreement for the static-potential to this order as well as the uniform spacing of $\frac{\pi}{R}$ of the excited levels.

22.2.2 A Possible Boundary Term

Higher order corrections to S_{eff} will typically involve coefficients with negative-integer mass dimensions. A mass dimension -1 term is only possible as a boundary term, which can certainly be there for the case of open-strings, while there can be no such terms for closed strings. One such boundary term, the simplest, considered by LW is

$$S_1 = \frac{b}{4} \int_0^T dz_0 \{ (\partial_1 h \cdot \partial_1 h)|_{z_1=0} + (\partial_1 h \cdot \partial_1 h)|_{z_1=R} \} \quad (22.10)$$

with b , a free parameter, having mass dimension -1. Power counting arguments would say this is a non-renormalizable interaction. But as LW point out in [3], there are no ultraviolet divergences. Presumably this is because this interaction is still quadratic. LW find corrections to the static-potential and energy level spacings to be

$$\Delta V(R) = -\frac{\pi(D-2)b}{24R^2} \quad \Delta E = \frac{\pi}{R} \left(1 + \frac{b}{R}\right) \quad (22.11)$$

Clearly such a term is not present in Eq. (22.1) as the quantization leading to this term had not included such boundary terms. Though for very large R this term is sub-leading to the Lüscher term, with accuracies that are able to probe even $\frac{1}{R^3}$ corrections, this would certainly show up unless b itself is very small. LW initially tried to account for their observed discrepancy of about 15% in the Lüscher term through such a boundary term with $b = 0$ in $D = 3$ and $b = 0.04 \text{ fm}$ in $D = 4$. Hari Dass and Pushan Majumdar, however found better fits without such boundary terms, and with only the universal $\frac{1}{R^3}$ terms of the truncated Arvis potential. In [3](section § 2.3) it is further pointed that at this order, the sole effect of this boundary interaction is the shift $R \rightarrow R - b$. In the $D = 4$ studies of [4,5], the lattice spacing was $\frac{1}{6} \text{ fm}$ and a b of 0.04 fm would amount to a noticeable shift of one quarter of lattice spacing. In the next section we discuss the arguments of LW based on what they called open-closed string duality to argue that b should actually be zero.

22.2.3 Dimension-2 Corrections

Motivated by several issues raised by their results in [1], LW recognized the need to study even higher order corrections to the spectrum of effective string theories. In their 2004 work [3] they undertook systematic studies of the same, and provided powerful calculational tools for the same. In particular, they focussed on potential (pun intended) corrections at $\frac{1}{R^3}$ level.

They argued that such and higher order corrections ought to be captured by interaction terms that are localized either on the world-sheet or on its boundary. Furthermore, they emphasized, in accordance with what is known in QFT's, that terms that are removable by field redefinitions can be dropped. We shall, however, have some comments later on about the locality aspects. Then they write down two possible interaction terms, both dimension 2, that are in addition to the dimension 1 boundary term already discussed. These are

$$\begin{aligned} S_2 &= \frac{c_2}{4} \int d^2z (\partial_a h \cdot \partial_a h)(\partial_b h \cdot \partial_b h) \\ S_3 &= \frac{c_3}{4} \int d^2z (\partial_a h \cdot \partial_b h)(\partial_a h \cdot \partial_b h) \end{aligned} \quad (22.12)$$

At this point LW introduce the notion of *Open-closed string duality* and draw many conclusions from it. In the present context, two of the important conclusions are (i) vanishing of the boundary term i.e. $b = 0$ in all dimensions, and, (ii) a linear constraint between c_2 and c_3 . We discuss these aspects in the next subsection.

22.2.4 Open-Closed String Duality

LW point out two equivalent interpretations of the Polyakov loop correlators. First is by considering the Polyakov lines to run along the time axis (with periodic boundary conditions in the time direction); this leads to the gauge theory interpretation of their correlation function as propagation in the presence of a $Q\bar{Q}$ -pair, and the subsequent effective string interpretation in terms of open strings. Equally well, the compact dimension with periodic boundary conditions could be taken along what was earlier thought to be a spatial direction. In this case, the effective string description would be that of a closed string propagation. This dual interpretation leads to relations between the partition functions in the two pictures.

In particular, for free string propagation (by this LW mean partition functions with only the leading S_0 term) for example, this leads to the relation

$$\mathcal{Z}_0 = \left(\frac{T}{2R}\right)^{\frac{(D-2)}{2}} \mathcal{Z}_0|_{T \leftrightarrow 2R} \quad (22.13)$$

To demonstrate this they make use of the elegant modular transformation properties of the Dedekind- η function:

$$\eta(q) = \left(\frac{2R}{T}\right)^{\frac{1}{2}} \eta(\tilde{q}) \quad \tilde{q} = e^{-\frac{4\pi R}{T}} \quad (22.14)$$

Let us first consider their analysis of the implications of this duality at the next higher order i.e. including the boundary term S_1 . Because of the quadratic nature of this correction, LW found a simple relation relating the partition function $\mathcal{Z}$ to this order and $\mathcal{Z}_0$:

$$\mathcal{Z} = (1 - \sigma T b) \mathcal{Z}_0 - b \frac{\partial}{\partial R} \mathcal{Z}_0 \quad (22.15)$$

i.e. the duality constraints on $\mathcal{Z}$ from those of $\mathcal{Z}_0$. One then expands the partition function for large T , and the dual for large R . Upon doing so, LW found that a non-vanishing b is inconsistent with duality and they drew the first important conclusion that $b = 0$.

Because of the absence of S_1 , standard path integral perturbation methods yield $\mathcal{Z}$ when the dimension 2 corrections S_2, S_3 are taken into account, with the result

$$\mathcal{Z} = \mathcal{Z}_0 \{1 - \langle S_2 \rangle_0 - \langle S_3 \rangle_0\} \quad (22.16)$$

where $\langle \dots \rangle_0$ denote expectation values in the free theory. The required expectation values can again be expressed in terms of $q, \eta(q)$ and its derivatives [3, 10]. We skip the details and go straight to the constraints imposed by duality for c_2, c_3 :

$$(D-2)c_2 + c_3 = \frac{D-4}{2\sigma} \quad (22.17)$$

Before going into the consequences of this, we mention the demonstration by LW (in appendix C of [3]) that the full Arvis spectrum mentioned before satisfies the

open-closed string duality exactly. Consequently, the truncated Arvis potential of Eq. (22.2) and energy levels of Eq. (22.5) also satisfy the duality constraints.

To appreciate their implications, let us look at the $\frac{1}{R^3}$ corrections that were worked out by LW in their Eqs. (4.9) and (4.10):

$$\begin{aligned}\Delta V(R) &= \left(\frac{\pi}{24}\right)^2 \frac{(D-2)}{2R^3} \{2c_2 + (D-1)c_3\} \\ \Delta E_1 &= \frac{\pi^2}{24R^3} \{(12D-14)c_2 + (5D+7)c_3\}\end{aligned}\quad (22.18)$$

Here only the results for the first excited state has been displayed, but it is more transparent to use their Eq. (6.1) for $n \leq 3$:

$$\Delta E_{n,i} = \frac{\pi^2}{R^3} \left\{ n \left[\frac{(D-2)}{12} - n \right] c_2 + \nu_{n,i} (c_3 + 2c_2) \right\} \quad (22.19)$$

where n refers to the level and i to its splittings. LW tabulate the coefficients $\nu_{n,i}$ in their Table 2, but we shall not be needing them.

At this point LW draw many conclusions from the duality constraints. An obvious one, according to them, is $c_3 = -2c_2$ in $D = 4$. We shall shortly argue that this is so for a large class of actions in *all* D . Equation (22.19) also reveals that when this condition is met, the terms with $\nu_{n,i}$ drop out completely. LW also point out that in $D = 3$ the integrands of the two actions are the same. Hence the spectrum is only sensitive to the combination $c_2 + c_3$. The duality constraint in this case reads $c_2 + c_3 = -\frac{1}{2\sigma}$. So they state that without loss of generality one can choose $c_3 = -2c_2$. This would then yield $c_2 = \frac{1}{\sigma}$ and $c_3 = -\frac{2}{\sigma}$.

While the author agrees with these values for other reasons, he has reservations about their arguments. Firstly, if the spectrum did depend on $c_2 + c_3$ only, whose value is completely fixed by the duality constraint in $D = 3$ to be $-\frac{1}{\sigma}$, nothing more needs to be done i.e. making the choice $c_3 = -2c_2$ is unwarranted. However, even in $D = 3$, Eq. (22.19) does not involve c_2, c_3 only in the combination $c_2 + c_3$, though it is true for their eqn(4.9) for $D = 3$. This can be seen explicitly by comparing the coefficients of c_2 and c_3 ;

$$c_2 : n \left[\frac{(D-2)}{12} - n \right] + 2\nu_{n,i} \quad c_3 : \nu_{n,i} \quad (22.20)$$

Their equality, for $D = 3$, is only possible if

$$\nu_{n,i} = n \left(n - \frac{1}{12} \right) \quad (22.21)$$

That is, $\nu_{n,i}$ should be *independent* of i . Their Table 2 seems to be at variance with this.

Irrespective of the actual values of $\nu_{n,i}$, as per their eqn(6.1), for given c_2 , there is a clear dependence on $c_3 + 2c_2$ and they can not set $c_3 + 2c_2 = 0$ (without loss, in

their words). Nevertheless, let us accept this and see what obtains. Firstly, the $\nu_{n,i}$ -dependence disappears. In $D = 3$, the duality constraint $c_2 + c_3 = -\frac{1}{2\sigma}$ uniquely fixes $c_2 = \frac{1}{\sigma}$ and $c_3 = -\frac{2}{\sigma}$. Consequently, the $D = 3$ spectrum to this order coincides with the truncated Arvis spectrum, in agreement with the results of [6,7] for the static potential.

Coming to $D = 4$, the duality constraint itself gives $c_3 + 2c_2 = 0$ and all dependence on $\nu_{n,i}$ vanishes, but either c_2 or c_3 remains undetermined. But [4,5] found agreement with truncated Arvis potential (they did not study the excited states; we will comment on the excited states later), with no room for any free parameters. The resolution of this will be discussed in the next section on the basis of the Polchinski-Strominger effective string theories. In that section, the issue of the legitimacy of using relations based on string theory like the Arvis potential, which should only be valid for $D = 26$, for arbitrary D values, will also be clarified. Before that, the author wishes to discuss a purely classical perspective on c_2, c_3 and the relationship between them.

22.2.5 Purely Classical Analysis

As c_2, c_3 are aspects of a classical action, one ought to be able to find any constraints on them in a purely classical manner without going into the details of the quantum spectrum, as LW did. Of course, in the end there has to be consistency between the classical and quantum approaches. Before delving into that, let us gain another perspective on the duality constraint between c_2 and c_3 , and for that let us consider those classical actions for which c_2 and c_3 do not explicitly depend on D . Then the duality constraint $(D - 2)c_2 + c_3 = \frac{D-4}{2\sigma}$ can only be satisfied for all D if and only if $c_3 + 2c_2 = 0$. This according to the author, is the primary significance of this relation though LW have discussed other interesting consequences of this relation. The immediate consequence of satisfying this is that $c_2 = \frac{1}{2\sigma}$ and $c_3 = -\frac{1}{\sigma}$, precisely the values that lead to the truncated Arvis spectrum. In what follows, all our remarks are for order $\frac{1}{R^3}$ only.

Are there such classical actions for which c_2, c_3 have these properties? We first examine the Nambu-Goto action itself as we have already argued that the dominant action for the LW analysis must be this, even though they did not explicitly state it. As discussed at length in Chap. 17, the Nambu-Goto action (we go back to the notations of Chap. 17) is

$$S_{NG} = -\mathcal{T} \int d\tau \int_0^\pi d\sigma \sqrt{\det(-g)} \quad g_{\alpha\beta} = \partial_\alpha X^\mu \partial_\beta X^\nu \eta_{\mu\nu} \quad (22.22)$$

with $\mathcal{T}$ being the string tension (σ in LW notation as also in [5], for example). Written out explicitly (see Chap. 17) [8, 13]

$$S_{NG} = -\mathcal{T} \int d\tau \int_0^\pi d\sigma \sqrt{\left(\frac{\partial X}{\partial \sigma} \cdot \frac{\partial X}{\partial \tau}\right)^2 - \left(\frac{\partial X}{\partial \sigma}\right)^2 \left(\frac{\partial X}{\partial \tau}\right)^2} \quad (22.23)$$

The most important aspects of the Nambu-Goto action for our purposes are the twin invariances of (i) *Target Space*-Lorentz invariance, and, (ii) world-sheet general coordinate invariance (sometimes fancily called diffeomorphism invariance). Both of them play crucial roles in what follows.

The general coordinate invariance allows the use of the *Static Gauge*: $X^0 = \tau$. In this gauge, the straight string configuration of a string fixed at the two end-points $\sigma = 0, \pi$ is given by $X^1 = \frac{R}{\pi} \sigma$, $X^i = 0$. The fluctuations about this are given by $X^i = h^i$ with the $D - 2$ transverse variables h^i obeying Dirichlet-boundary conditions. Thus the h^i can be identified with LW variables. The minimal surface is rectangular with extent T along X^0 direction and R along with X^1 -direction with area RT , contributing $\mathcal{T} RT$ to the action.

The $(-g)$ factor is easily worked out for the configuration with fluctuations. The result is

$$\det(-g) = (\partial_\sigma h^i \partial_\tau h^i)^2 - (-1 + (\partial_\tau h^i)^2) \left(\frac{R^2}{\pi^2} + (\partial_\sigma h^i)^2 \right) \quad (22.24)$$

Let us collect terms upto quadratic order in fluctuations to fix various scalings and normalizations:

$$\det(-g)_2 = \frac{R^2}{\pi^2} + (\partial_\sigma h^i)^2 - (\partial_\tau h^i)^2 \cdot \frac{R^2}{\pi^2} \quad (22.25)$$

This immediately suggests the rescaling $\sigma \rightarrow \sigma' = \frac{R}{\pi} \sigma$ so that the range of σ' is $(0, R)$ as in the LW conventions. After this scaling the quadratic part of the action becomes

$$S_0 = \frac{\mathcal{T}}{2} \int d\tau \int_0^R d\sigma' (\partial_a h)(\partial_a h) \quad (22.26)$$

So a rescaling of the h field by a $\sqrt{\mathcal{T}}$ makes the h -field dimensionless and yields precisely the S_0 of LW. Now we turn to the next higher order correction where the scalings done above are already incorporated. To avoid clutter, the $\prime$ on the rescaled σ will not be shown explicitly:

$$\mathcal{L}_2 = \frac{1}{2\mathcal{T}} \{ (\partial_\sigma h^i \partial_\tau h^i)^2 - (\partial_\tau h^i)^2 (\partial_\sigma h^j)^2 \} - \frac{1}{8\mathcal{T}} \{ (\partial_\sigma h^i)^2 - (\partial_\tau h^i)^2 \}^2 \quad (22.27)$$

The first term, after some algebra, can be reexpressed as

$$- \frac{1}{2\mathcal{T}} \frac{1}{2} \{ (\partial_a h \partial_b h)(\partial_a h \partial_b h) - (\partial_a h \partial_a h)(\partial_b h \partial_b h) \} \quad (22.28)$$

Combining all terms, one finally arrives at

$$S_2 + S_3 = \int d\tau \int_0^R d\sigma \left\{ \frac{c_2}{4} (\partial_a h \partial_a h)(\partial_b h \partial_b h) + \frac{c_3}{4} (\partial_a h \partial_b h)(\partial_a h \partial_b h) \right\} \quad (22.29)$$

with $c_2 = \frac{1}{2T}$ and $c_3 = -\frac{1}{T}$. This is precisely the form of the dim 2 terms written down by LW.

Thus we have deduced all the relationships between c_2 and c_3 ascribed by LW to string duality, for all values of D , by purely classical analysis. Not surprisingly, they also reproduce the truncated Arvis spectrum. Therefore the particular values of c_2, c_3 already saturate the $\frac{1}{R^3}$ terms seen in the numerical simulations of Hari Dass and Pushan Majumdar. Does it mean that the effective string theory has only the Nambu-Goto action? There is a priori no reason for that and even higher order terms with the same symmetry content as Nambu-Goto action are certainly possible, and will in fact be explicitly constructed later in the chapter. This only means that there are no further corrections to the spectrum at the $\frac{1}{R^3}$ level to what is given by the Nambu-Goto action. We will establish these results in the Polchinski-Strominger effective string theory. This is done in the next section where it will also be shown that potentially there can be order $\frac{1}{R^3}$ corrections to the spectrum even if there are no candidate actions at this level over and above the PS action.,

It is clear that in the static gauge, even just the Nambu-Goto action leads to an infinite number of higher order terms. This is a nuisance, as in each order in powers of $\frac{1}{R}$, one has to disentangle the effects of the Nambu-Goto term to understand genuinely new additions to the actions. As in most perturbation theory calculations, this quickly becomes very unwieldy. We will see in the next section that in the covariant gauge characterising Polchinski-Strominger(PS) theory [14], the Nambu-Goto action only produces the leading quadratic term.

22.3 Polchinski-Strominger(PS) Effective String Theory

Introducing the composite operator

$$g_{\alpha\beta} = \partial_\alpha X^\mu \partial_\beta X_\mu \quad (22.30)$$

with ξ^α being the coordinates of the world sheet, it is easy to see that under general coordinate transformations of the the world-sheet coordinates ξ^α , $g_{\alpha\beta}$ indeed transforms as a metric tensor provided X^μ transform as scalar fields. We call it a *metric substitute* and in this particular case, it has the meaning as the *induced metric* on the world-sheet. Thus the Nambu-Goto action Eq. (22.22) is indeed invariant under the world-sheet general coordinate transformations.

Just as in the case of the static gauge, this invariance also permits a *covariant gauge*, exactly as in the case of string theory (for details please see Chap. 17). Taking the world-sheet coordinates to be (τ, σ) and introducing the light-cone coordinates $\tau^\pm = \tau \pm \sigma$, this covariant gauge, also called the conformal gauge, is fixed by the conditions

$$g_{++}(\tau^\pm) = g_{--}(\tau^\pm) = 0 \quad (22.31)$$

The only remaining component of the metric substitute is $g_{+-} = g_{-+} = \partial_+ X \cdot \partial_- X$. The Nambu-Goto action then takes the simple looking form (we have chosen to

express it the same way as PS have; the string tension is $\frac{1}{4\pi a^2}$, with a having dimensions of length):

$$S_0^{PS} = \frac{1}{4\pi a^2} \int d\tau^+ d\tau^- \partial_+ X^\mu \partial_- X_\mu \quad (22.32)$$

This is the amazing property of the conformal gauge: the Nambu-Goto action is equivalent to a single, quadratic action, unlike the static gauge where it produced an infinite series of actions of increasing dimensions. In Chap. 17, we had called this the Nielsen-Susskind action, and also noted how Nambu had preferred the area action as being geometrical. But now we see that what he had rejected is just as geometrical as what he had preferred, one being just a gauge-fixed version of the other.

22.3.1 Leading Order Analysis of PS Effective Actions

Now we will analyse this leading order effective action, both to establish the general methodology as well as to rederive the Arvis spectrum in the conformal gauge. While this follows Polchinski and Strominger [14], the spirit is somewhat different as we shall not restrict to a perturbative expansion in $\frac{1}{R}$; instead we shall work out the exact spectrum, to all orders.

The equations of motion (Euler-Lagrange)(EOM) following from this action are:

$$\partial_+ \partial_- X^\mu = 0 \quad (22.33)$$

The classical ground state, analogous to the rigid string solution of LW is expressed as

$$X_{cl}^\mu(\tau^\pm) = e_+^\mu R \tau^+ + e_-^\mu R \tau^- \quad (22.34)$$

where $e_\pm^\mu$ are constant vectors. Actually, the most general solutions of the EOM are

$$X^\mu(\tau^\pm) = F^\mu(\tau^+) + G^\mu(\tau^-) \quad (22.35)$$

But S_0^{PS} , even after the gauge fixing, still has *residual* local invariances under $\tau^+ \rightarrow \tau^+ + \epsilon^+(\tau^+)$ and $\tau^- \rightarrow \tau^- + \epsilon^-(\tau^-)$, with the corresponding transformations of X^μ :

$$\delta_+ X^\mu = \epsilon^+(\tau^+) \partial_+ X^\mu \quad \delta_- X^\mu = \epsilon^-(\tau^-) \partial_- X^\mu \quad (22.36)$$

This allows F^μ, G^μ to be chosen as in Eq. (22.35).

Following PS, we also choose to analyse closed strings only though the numerical determinations were done for open strings. However, as LW have discussed, there is a duality between them. Also, we need not consider issues concerning boundary terms in the action. Thus, all solutions, and the classical ones in particular, satisfy the periodicity conditions

$$X^\mu(\tau, \sigma + 2\pi) = X^\mu(\tau, \sigma) + 2\pi R \delta_1^\mu \quad (22.37)$$

where the string is taken to lie along the X^1 -direction. This immediately leads to $e_+^\mu - e_-^\mu = \delta_1^\mu$. The classical constraints of Eq. (22.31) yield $e_+ \cdot e_+ = e_- \cdot e_- = 0$ i.e. these are null vectors. Combining this with the periodicity condition yields $e_+ \cdot e_- = -\frac{1}{2}$.

The conserved energy-momentum tensors are

$$T_{--} = -\frac{1}{2a^2} \partial_- X \cdot \partial_- X \quad T_{++} = -\frac{1}{2a^2} \partial_+ X \cdot \partial_+ X \quad (22.38)$$

The fluctuations are now characterized by Y^μ given by $X^\mu = X_{cl}^\mu + Y^\mu$. In terms of these, T_{--} takes the form (analogous expression for T_{++} with $e_- \rightarrow e_+$, $\partial_- \rightarrow \partial_+$):

$$T_{--} = -\frac{R}{a^2} e_- \cdot \partial_- Y - \frac{1}{2a^2} \partial_+ Y \cdot \partial_+ Y \quad (22.39)$$

and likewise for T_{++} . Next, one introduces the mode expansions

$$\partial_- Y = a \sum_{m=-\infty}^{\infty} \alpha_m^\mu e^{-im\tau^-} \quad \partial_+ Y = a \sum_{m=-\infty}^{\infty} \tilde{\alpha}_m^\mu e^{-im\tau^+} \quad (22.40)$$

The oscillator algebra is given by

$$[\alpha_m^\mu, \alpha_n^\nu] = m \eta^{\mu\nu} \delta_{m+n,0} \quad [\tilde{\alpha}_m^\mu, \tilde{\alpha}_n^\nu] = m \eta^{\mu\nu} \delta_{m+n,0} \quad (22.41)$$

These are the same as what was encountered for free closed strings in Chap. 17 and for the Shapiro-Virasoro(SV) model in Chap. 16. The Virasoro generators are again defined in terms of Fourier expansions of T_{--} and T_{++} . The result is

$$L_n = \frac{R}{a} e_- \cdot \alpha_n + \frac{1}{2} \sum_{m=-\infty}^{\infty} : \alpha_{n-m} \cdot \alpha_m : \quad (22.42)$$

As in the earlier discussion of string theories, $: \dots :$ denotes normal ordering. Likewise for $\tilde{L}_n$ with $e_- \rightarrow e_+$ and $\alpha_n \rightarrow \tilde{\alpha}_n$. It is easy to verify the Virasoro algebras in both sectors to be

$$[L_m, L_n] = (m-n) L_{m+n} + \frac{D}{12} (m^3 - m) \delta_{m+n,0} \quad (22.43)$$

likewise for $\tilde{L}_n$. The additional R -dependent term in the Virasoro generators does not contribute to the central charge by virtue of $e_- \cdot e_- = e_+ \cdot e_+ = 0$. All these manipulations are exactly as in string theories as discussed in Chap. 17. Therefore, the critical dimension remains 26.

As before, the quantum ground state is $|k, k; 0\rangle$ which is an eigenstate of both $\alpha_0^\mu, \tilde{\alpha}_0^\mu$ with the same eigenvalue $a k^\mu$. The physical state conditions are $L_0 = \tilde{L}_0 = 1$ and $L_n = \tilde{L}_n = 0$, for $n \geq 2$. By taking sum and difference of the $L_0, \tilde{L}_0$ con-

ditions, and using the eigenvalue conditions, along with the periodicity condition $e_+^\mu - e_-^\mu = \delta_1^\mu$, it is easy to arrive at

$$k^1 = 0 \quad k^2 + \frac{R}{a^2}(e_+ + e_-) \cdot k = \frac{2}{a^2} \quad (22.44)$$

The total momentum of the string is (see section § 4.2 later)

$$p^\mu = \frac{R}{2a^2}(e_+^\mu + e_-^\mu) + \frac{1}{2a}(\alpha_0^\mu + \tilde{\alpha}_0^\mu) \quad (22.45)$$

which for the ground state is given by $p_{grnd}^\mu = \frac{R}{2a^2}(e_+ + e_-)^\mu + k^\mu$. Therefore, the total rest energy of the ground state is

$$(-p^2)^{\frac{1}{2}} = \sqrt{\left(\frac{R}{2a^2}\right)^2 - k^2 - \frac{R}{a^2}(e_+ + e_-) \cdot k} \quad (22.46)$$

which, on using the second of Eq. (22.44) becomes

$$(-p^2)^{\frac{1}{2}}|_{grnd} = \frac{R}{2a^2} \sqrt{1 - 2\left(\frac{2a}{R}\right)^2} \quad (22.47)$$

which is the Arvis result for closed strings, but for $D = 26$! The point is that even though Eq. (22.3) is made to look like a function of D , it is only meaningful in $D = 26$, as demonstrated by Arvis himself on requiring consistency with rotational invariance (see Chap. 17 for the details). Therefore, it is only legitimate to use the Arvis potential for $D = 26$.

The big question then is the legitimacy of using the expression for the Lüscher term and the truncated Arvis potential for $D \neq 26$. PS give a resolution of this and we discuss it in the next section.

22.4 PS Effective String Theories for all D

Polchinski and Strominger [14] constructed effective string theories that are valid in all dimensions D , not just $D = 26$. Let us recall the origin of the $D = 26$ difficulty in string theories (see chapter 17 for details). One way of understanding this is through the *Conformal Anomaly* (for a brief discussion of anomalies the reader is referred to Chap. 18). Only in 26 dimensions is this anomaly absent. However, Polyakov [15] had shown that strings can be formulated in $D < 26$ dimensions. This is called Polyakov subcritical string theory.

As an alternative to the Nambu-Goto action, Polyakov introduces an action, by enlarging the field content to also include an *intrinsic metric* $h_{\alpha\beta}$ on the world sheet transforming exactly as the induced metric of Eq. (22.30) under world-sheet general coordinate transformations:

$$S_{polya} = \int d^2\xi \sqrt{-h} h^{\alpha\beta} \partial_\alpha X \cdot \partial_\beta X \quad (22.48)$$

where h is the determinant of $h^{\alpha\beta}$. With the inclusion of the intrinsic metric, invariance under world-sheet general coordinate invariance now becomes exactly akin to that in General Relativity theory, unlike that in the case of the Nambu-Goto action, as already remarked. However, the Polyakov action is also invariant under *local Weyl-scalings*:

$$h_{\alpha\beta} \rightarrow \omega(\xi) h_{\alpha\beta} \quad X^\mu \rightarrow X^\mu \quad (22.49)$$

This forms the symmetry content of the string theory now. Polyakov analysed this theory in the path-integral approach. We will have lot more to say about this class of actions shortly.

Fujikawa had shown [16], in the context of chiral anomalies, that in the path-integral formulation, the entire anomaly resides in the path-integral measure. Now in $D \neq 26$, the Weyl-anomaly also resides in the path-integral measures for the functional integrations over X^μ and $h_{\alpha\beta}$. By utilizing general coordinate transformations, every two dimensional metric can be brought into a *conformally flat* metric and the only degree of freedom left is the conformal factor $e^\phi(\xi)$. Now the measure for ϕ -integrations being no longer invariant, these integrations induce an action for the additional degree of freedom ϕ . We refer the reader to [15] for this very deep aspect and simply quote the result for this so called *Liouville Action*:

$$S_L = \frac{26-D}{48\pi} \int d^2\xi \partial_+ \phi \partial_- \phi \quad (22.50)$$

The way to understand this beautiful work of Polyakov is to realise that the additional degree of freedom makes the right additional contribution to the central charge of the Virasoro algebra so as to completely cancel the conformal anomaly in every D. Viewed as a fundamental theory, the Liouville theory also has some problems like the so called $c = 1$ barrier, but as an effective theory, hopefully, this will not be a problem. Actually, the Liouville action also has a difficult to handle, $e^{\mu\phi}$ part, which PS have ignored. For the sake of formulating an effective theory, as PS do, this omission is permissible. Not surprisingly, this action is proportional to $(26-D)$. Without the exponential term, the action is invariant under shifts of the ϕ -field. Now Polchinski and Strominger introduce an effective action only for X^μ by equating h_{+-} of the conformal gauge to the induced metric g_{+-} introduced earlier i.e by choosing $e^\phi = g_{+-}$. Calling this L , the effective action one obtains is

$$S_{PL} = \frac{26-D}{48\pi} \int d\tau^+ d\tau^- \frac{\partial_+ L \partial_- L}{L^2} \quad (22.51)$$

exactly invariant under the same conformal transformations of Eq. (22.36) that left S_0^{PS} invariant. This action is clearly non-polynomial in general. As configurations can easily be found for which the denominator can vanish i.e. $\partial_+ X \cdot \partial_- X = 0$, it is also singular in general. But as an effective action for *long* strings whose classical configuration satisfies $L = -\frac{R^2}{2}$, it should be acceptable (unless proven otherwise).

Instead of analysing their effective action in this beautiful form, PS now proceed to make things superficially confusing by throwing away all terms proportional to the

leading Euler-Lagrange equations of motion(EOM) from S_0^{PS} i.e. $\partial_+ \partial_- X^\mu = 0$. They do this by suitable *field redefinitions* of X^μ resulting in the somewhat simpler action

$$S_{PS} = \frac{\beta}{4\pi} \int d\tau^+ d\tau^- \frac{\partial_+^2 X \cdot \partial_- X \partial_-^2 X \cdot \partial_+ X}{(\partial_+ X \cdot \partial_- X)^2} \quad (22.52)$$

In addition to throwing away all terms proportional to the leading order EOM, they also throw caway all terms proportional to the leading order constraints $\partial_+ X \cdot \partial_+ X = 0 = \partial_- X \cdot \partial_- X$ (see [14] for details). They also kept the coefficient of the action as a free parameter $\frac{\beta}{4\pi}$ to be determined later.

But the price to be paid for this field redefinition is that the transformation laws under which the simplified action(plus S_0^{PS}) are invariant looks like a real mess! PS had to determine this iteratively i.e. by truncating the action to a certain order in $\frac{1}{R}$, the transformation laws were also determined to a certain suitable order, a decidedly tedious procedure. Neglecting terms of order $\frac{1}{R^3}$, equivalently keeping only terms upto order $\frac{1}{R^2}(O(R^{-3}))$, they determined the transformation laws

$$\delta X^\mu = \epsilon^-(\tau^-) \partial_- X^\mu - \frac{\beta a^2}{2} \partial_-^2 \epsilon^-(\tau^-) \frac{\partial_+ X^\mu}{\partial_+ X \cdot \partial_- X} \quad (22.53)$$

It is to be appreciated that the orders of truncations in the action and the transformation laws should be self-consistent. It is an elementary exercise to work out that if terms of order R^{-N} are kept in the action, terms to the same order must be retained in the transformation laws also. In their original analysis of [14], they retained terms of order R^{-2} in both of them(i.e. $O(R^{-3})$). Obviously, the closure of the algebra of transformations will also be valid only upto some order.

When we started working on the PS theory,¹ we took on face value their statement, immediately after their Eq.(5) where they had stated the Liouville action of Eq. (22.50), that substitution of the induced metric in place of e^ϕ gives their action of Eq. (22.52). It is their transformation law that baffled this author. The clue finally came on his working out the algebra of their transformation laws of Eq. (22.53) to be

$$[\delta_-^{PS}(\epsilon_1^-), \delta_-^{PS}(\epsilon_2^-)] = \delta_-^{PS}(\epsilon_{12}^-) + O(R^{-4}) \quad (22.54)$$

with $\epsilon_{12}^- = \epsilon_1^- \partial_- \epsilon_2^- - \epsilon_2^- \partial_- \epsilon_1^-$. The algebra of the conformal transformations of Eq. (22.36) is

$$[\delta_-^0(\epsilon_1^-), \delta_-^0(\epsilon_2^-)] = \delta_-^0(\epsilon_{12}^-) \quad (22.55)$$

which is exact i.e. valid to *all* orders in R . This meant that the PS transformation laws were just conformal transformations in disguise! This also meant that the strange-looking PS transformations and conformal transformations ought to be related by

¹ This author became aware of the PS approach from remarks by Professor Julius Kuti as conveyed by Professor Apoorva Patel.

field redefinitions, to some order of accuracy. The author then by hand constructed such a field redefinition and was surprised that the Liouville action could be recovered. Furthermore, it could be seen that the Liouville action itself was *exactly* invariant under Eq. (22.36), an important point that had not been stressed sufficiently by PS.

Now we present the original analysis of PS of their effective action which removes the conformal anomaly in every D. It centers around the modifications to the Virasoro generators and their algebra due to the added effective action. For this, one needs to construct the conserved energy-momentum tensors T_{--} , T_{++} through Noether's theorem though care has to be exercised as conformal invariance is local. The new *On-shell* energy-momentum tensors are (i.e. after throwing away terms proportional to the EOM's):

$$T_{--}^{PS} = -\frac{1}{2a^2} \partial_- X \cdot \partial_- X + \frac{\beta}{2L^2} (L \partial_-^2 L - (\partial_- L)^2 + \partial_- X \cdot \partial_- X \partial_+^2 X \cdot \partial_-^2 X - \partial_+ L \partial_- X \cdot \partial_-^2 X) \quad (22.56)$$

It is a bit tedious to compare this with what Drummond has given in his Eq. (2.12) [17]. We will soon compare our corresponding expressions for fluctuations to order R^{-2} (the relevant order to compute the spectrum to order R^{-3}). As the leading term in the transformation law grows as R , the leading term of the energy momentum tensor also grows like R , and as the action is $O(R^{-3})$, one can only get $T_{\pm\pm}$ to $O(R^{-2})$. In terms of the fluctuations Y^μ introduced earlier, the energy momentum tensors now read (this should be compared with the result of the leading order analysis in Eq. (22.39)):

$$T_{--} = -\frac{R}{a^2} e_- \cdot \partial_- Y - \frac{1}{2a^2} \partial_- Y \cdot \partial_- Y - \frac{\beta}{R} e_+ \cdot \partial_-^3 Y + \dots \quad (22.57)$$

It is clear, just by inspection, that the interplay between the first and last terms has the effect of shifting the central charge by 12β i.e. from D to $D + 12\beta$! PS show this by working out the *Operator Product Expansion* (OPE) for T_{--} , T_{++} .

It can equally well be shown by constructing the Virasoro generators L_n by Fourier-analysing the energy momentum tensors. Following the standard methods discussed earlier, the generators are found to be

$$L_n = \frac{R}{a} e_- \cdot \alpha_n + \frac{1}{2} \sum_{m=-\infty}^{\infty} : \alpha_{n-m} \cdot \alpha_m : + \frac{\beta}{2} \delta_{n,0} - \frac{a\beta n^2}{R} e_+ \cdot \alpha_n + O(R^{-2}) \quad (22.58)$$

and likewise for $\tilde{L}_n$ in terms of $\tilde{\alpha}_n$. It is again a straightforward exercise to show that they satisfy the algebra

$$[L_m, L_n] = (m - n) L_{m+n} + \frac{D + 12\beta}{12} (m^3 - m) \delta_{m+n,0} \quad (22.59)$$

likewise for $\tilde{L}_n$. The additional contributions to the central extension term come from two important sources: (a) the cross terms from the first and last terms in L_n where

the factors of R exactly compensate each other, and, (b) the shift of L_0 by $\frac{\beta}{2}$. As already discussed in Chap. 16, the general structure of this term follows from very general considerations like Jacobi identities, only the coefficient needs some explicit calculations.

The implications of this modification for the criticality of strings is dramatic. Instead of the earlier requirement of $D = 26$, the new requirement is $D + 12\beta = 26$ which can be satisfied for every D as long as β is chosen to be the critical value $\beta_c = -\frac{D-26}{12}$. Remarkably, the coefficient $\frac{\beta}{4\pi}$ of the action takes the same value as in the original Polyakov-Liouville theory. This leads to the final expression for the Virasoro generators:

$$L_n = \frac{R}{a} e_- \cdot \alpha_n + \frac{1}{2} \sum_{m=-\infty}^{\infty} : \alpha_{n-m} \cdot \alpha_m : + \frac{\beta_c}{2} \delta_{n,0} - \frac{a\beta_c n^2}{R} e_+ \cdot \alpha_n + O(R^{-2}) \quad (22.60)$$

In what follows we shall always mean β_c even if we use β .

We now analyse the ground state energy; the case of excited states is straightforward to generalize. The quantum ground state $|k, k; 0\rangle$ is a simultaneous eigenstate of both α_0^μ and $\tilde{\alpha}_0^\mu$ with the eigenvalue $a k^\mu$. The sum and difference of the conditions $L_0 = \tilde{L}_0 = 1$ now yields

$$k^1 = 0 \quad k^2 + \frac{R}{a^2} (e_+ + e_-) \cdot k = \frac{2 - \beta_c}{a^2} \quad (22.61)$$

To this order, the ground state momentum still remains $p_{grnd}^\mu = \frac{R}{2a^2} (e_+^\mu + e_-^\mu) + k^\mu$ (this will be explained in detail shortly). The resulting ground state mass becomes

$$(-p^2)^{\frac{1}{2}}|_{grnd} = \frac{R}{2a^2} \sqrt{1 - \frac{D-2}{12} \left(\frac{2a}{R}\right)^2} \quad (22.62)$$

But this analysis is only valid to $O(R^{-2})$ and therefore self-consistency requires that this be truncated to this order, yielding

$$V(R) = \frac{R}{2a^2} - \frac{D-2}{12} \frac{1}{R} + \dots \quad (22.63)$$

giving the universal Lüscher term for closed strings, but now for all D . This vindicates the use of the Lüscher term in interpreting numerical simulation results for $D \neq 26$. Now we turn to showing how the truncated Arvis potential, valid for all D , also emerges from the PS theory at the next non-trivial higher order.

22.4.1 Order R^{-3} Corrections to the Spectrum

Now we address the important issue of possible terms in the spectrum at order R^{-3} i.e. $O(R^{-4})$. The truncated Arvis potential to which we had claimed numerical

evidence in $D = 3$ SU(2) and $D = 4$ SU(3) [4–7] is of this type. The truncated Arvis spectrum is the analogous result for excited states. To apply the PS formalism to this end requires going beyond their analysis. Following their methodology would require on the one hand identification of possible additional terms to the action, and on the other hand find the modifications to their transformation laws that would keep the total action invariant. PS themselves stated, without any elaboration, that the next such terms in the action are of order R^{-4} . Since everything depends on the correctness of this assertion, a more systematic analysis is called for. There is also the related issue of extending the transformation laws.

This problem was independently solved by Drummond in [17], and by this author and Peter Matlock in their work [18]. Both the approaches were technically almost identical. While Hari Dass and Matlock had shown that there were no additional terms in the action at R^{-3} order over and above what one obtains on expanding the PS-action to that level, they had also claimed that there would be such additions at R^{-4} and R^{-5} orders, Drummond had claimed that in fact the possible corrections would only be at order R^{-6} . Hari Dass and Matlock had criticised the claims made by Drummond regarding terms of order higher than R^{-3} , and in response Drummond [19] had given a more explicit and systematic treatment which upheld his earlier claims. So indeed the potential corrections would, rather remarkably, only appear at order R^{-6} and higher. Shortly afterwards, Hari Dass and Matlock gave a systematic method for constructing all action candidates that had the virtue of keeping the transformation laws fixed as those that left the leading action S_0^{PS} invariant. In other words, their construction, which they called a *Covariant Calculus for Effective String Theories* [20,21] enables one to construct actions that are invariant to all orders in R^{-1} under the same fixed transformation laws as given by Eq. (22.36), without the need to keep adjusting the transformation laws every time a new term was added to the action as per the methods of PS. We will discuss the covariant calculus in detail shortly. The upshot of that calculus was a straightforward vindication of Drummond's claims about the only corrections being at order R^{-6} and higher, but of the four candidate actions proposed by Drummond, only two specific combinations appeared.

PS themselves had given a recipe for constructing the additional terms to their effective action. In essence, it involved constructing actions that transform like (1, 1) in a naive sense meaning the net number of + and – indices are one each. This naive criterion should be distinguished from quantities actually transforming as (1, 1) under transformations analogous to the PS-transformation laws, but the complication now is that the forms of these transformation laws actually depend on the precise choice of terms in the action. The naive criterion is necessary but not sufficient. In fact, the PS lagrangean is a case in point, it is only (1, 1) in the naive sense and does not strictly transform as a (1, 1) tensor; nevertheless, the PS-action is invariant. This distinction will be made clearer when we discuss the covariant calculus.

The typical form of such terms is one with a numerator with certain number of (+, –) indices and a denominator with the correct numbers of these indices to provide a net (1,1) term. The denominators should not become singular even in the limited

context of fluctuations around a classical background of a rigid string. That only leaves powers of $L = \partial_+ X \cdot \partial_- X$. This is certainly so for the PS-lagrangian.

In the next part of the recipe, PS advocate throwing away all terms proportional to the leading order EOM's as they can be eliminated through field redefinitions albeit at the cost of changing the transformation laws. Finally, PS also advocate throwing away terms proportional to the leading order constraints $\partial_\pm X \cdot \partial_\pm X$. Needless to say, this recipe gets rapidly unwieldy with each increasing order.

At first, both Drummond, and, Hari Dass and Matlock(HM) followed this procedure. Initially, the latter had claimed potential corrections at order R^{-4} , R^{-5} , but Drummond in [19] showed that all these are actually equivalent to order R^{-6} terms after partial integrations. We refer the reader to these sources for details but the important punchline agreed to by both the groups is *the absence of terms at order R^{-3} over and above what one already obtained from expanding the PS-action to this order*. HM also showed that the PS-transformation laws are actually good to terms including R^{-3} terms.

The consequence of this is that the on-shell energy-momentum tensors can be computed to include terms of order R^{-2} by simply expanding Eq. (22.56). The result (with a similar expression with $+$ $\leftrightarrow$ $-$) is

$$T_{--} = -\frac{R}{a^2} e_- \cdot \partial_- Y - \frac{1}{2a^2} \partial_- Y \cdot \partial_- Y - \frac{\beta}{R} e_+ \cdot \partial_-^3 Y - \frac{\beta}{R^2} \{2(e_+ \cdot \partial_-^2 Y)^2 + 2e_+ \cdot \partial_-^3 Y(e_+ \cdot \partial_- Y + e_- \cdot \partial_+ Y) + 2e_- \cdot \partial_-^2 Y e_- \cdot \partial_+^2 Y + \partial_+ Y \cdot \partial_-^3 Y\} \quad (22.64)$$

This is in complete agreement with that of Drummond as in his Eq. (2.16). However, he uses somewhat different techniques in his next steps. We shall continue to follow [18].

This energy-momentum is conserved on-shell i.e.

$$\partial_+ T_{--} = 0 \quad (22.65)$$

While upto now T_{--} only involved derivatives of Y^μ wrt τ^- and this conservation was obvious in view of the EOM $\partial_+ \partial_- Y = 0$. But now, at order R^{-2} , there are two terms that involve derivatives of Y wrt τ^+ and it is not immediately obvious how the conservation law follows. For this we need to examine the EOM to order R^{-3} . We quote the result for this EOM and refer the interested reader to [18]:

$$\begin{aligned} \frac{2}{a^2} \partial_{+-} Y^\mu = & -4 \frac{\beta}{R^2} e_+ \cdot \partial_+^2 \partial_-^2 Y e_-^\mu - \frac{4\beta}{R^3} \{\partial_+^2 [\partial_-^2 Y^\mu (e_+ \cdot \partial_- Y + e_- \cdot \partial_+ Y)] \\ & + \partial_-^2 [\partial_+^2 Y^\mu (e_+ \cdot \partial_- Y + e_- \cdot \partial_+ Y)] + 4 e_+^\mu \partial_- (\partial_+^2 Y \cdot \partial_-^2 Y) + 4 e_-^\mu \partial_+ (\partial_+^2 Y \cdot \partial_-^2 Y)\} \end{aligned} \quad (22.66)$$

This can be simplified considerably on noting that all terms involving $\partial_+ \partial_- Y$ and its derivatives will actually be higher order in R^{-1} than relevant. The simplified

form of EOM becomes

$$\begin{aligned} \frac{2}{a^2} \partial_{+-} Y^\mu &= \frac{4\beta}{R^3} \{ e_-^\mu \partial_+ (\partial_+^2 Y \cdot \partial_-^2 Y) + e_+^\mu \partial_- (\partial_+^2 Y \cdot \partial_-^2 Y) \\ &\quad - \partial_-^2 (\partial_+^2 Y^\mu e_+ \cdot \partial_- Y) - \partial_+^2 (\partial_-^2 Y^\mu e_- \cdot \partial_+ Y) \} \end{aligned} \quad (22.67)$$

This can be solved iteratively by writing $Y^\mu = Y_0^\mu + Y_1^\mu$ with Y_0^μ being the solution to the leading order EOM, leading to

$$\begin{aligned} \frac{2}{a^2} \partial_{+-} Y_1^\mu &= \frac{4\beta}{R^3} \{ e_-^\mu \partial_+ (\partial_+^2 Y_0 \cdot \partial_-^2 Y_0) + e_+^\mu \partial_- (\partial_+^2 Y_0 \cdot \partial_-^2 Y_0) \\ &\quad - \partial_-^2 (\partial_+^2 Y_0^\mu e_+ \cdot \partial_- Y_0) - \partial_+^2 (\partial_-^2 Y_0^\mu e_- \cdot \partial_+ Y_0) \} \end{aligned} \quad (22.68)$$

This can be readily integrated to give

$$\frac{2}{a^2} \partial_- Y^\mu = \frac{4\beta}{R^3} \{ e_+^\mu \partial_+ Y_0 \cdot \partial_-^2 Y_0 + e_-^\mu \partial_-^2 Y_0 \cdot \partial_+^2 Y_0 - \partial_-^2 Y_0^\mu e_- \cdot \partial_+^2 Y_0 - \partial_+^2 Y_0^\mu e_+ \cdot \partial_-^2 Y_0 \} \quad (22.69)$$

On substituting this into the first term $-\frac{R}{a^2} e_- \cdot \partial_- Y$ of T_{--} , one sees the remarkable cancellation of all terms with $+$ -derivatives, leaving behind an expression for T_{--} that is only built with $-$ -derivatives of Y_0 , satisfying the conservation law in a straightforward manner. This is the purely holomorphic representation of T_{--} (in [18] there was a typo and we had used *holonomic*, a meaningless phrase in this context!). Now it is also clear why we had to solve EOM to order R^{-3} to get this cancellation to work at order R^{-2} . The final expression for T_{--} is:

$$\begin{aligned} T_{--} &= -\frac{R}{a^2} e_- \cdot \partial_- Y_0 - \frac{1}{2a^2} \partial_- Y_0 \cdot \partial_- Y_0 - \frac{\beta}{R} e_+ \cdot \partial_-^3 Y_0 \\ &\quad - \frac{2\beta}{R^2} e_+ \cdot \partial_-^3 Y_0 e_+ \cdot \partial_- Y_0 - \frac{2\beta}{R^2} (e_+ \cdot \partial_-^2 Y_0)^2 \end{aligned} \quad (22.70)$$

This fully agrees with the Drummond's expression. With the mode expansion for Y_0^μ already discussed, the Virasoro generators at this order are:

$$\begin{aligned} L_n &= \frac{R}{a} e_- \cdot \alpha_n + \frac{1}{2} \sum_{m=-\infty}^{\infty} : \alpha_{n-m} \alpha_m : + \frac{\beta_c}{2} \delta_{n,0} - \frac{a\beta_c n^2}{R} e_+ \cdot \alpha_n \\ &\quad - \frac{\beta_c a^2 n^2}{R^2} \sum_{m=-\infty}^{\infty} : e_+ \cdot \alpha_{n-m} e_+ \cdot \alpha_m : + O(R^{-3}) \end{aligned} \quad (22.71)$$

Agreeing with Drummond's formulae for the same.

Now, the additional modifications to $L_0, \tilde{L}_0$ actually vanish! Thus the spectrum of PS-effective string theories to order R^{-3} is the same as that of the truncated Arvis potential, but now with the dramatic reinterpretation that it is valid for all D, and not just $D = 26$!

22.4.2 Ground State Momentum Revisited

There is, however, a serious caveat to the above derivation. It depends on the tacit assumption that the ground state momentum does not receive any corrections even at R^{-3} order and the earlier expression $p_{grnd}^\mu = \frac{R}{2a^2}(e_+^\mu + e_-^\mu) + k^\mu$ still holds. Drummond just states this without elaboration. As this is a crucial ingredient and also as the derivation of this has a lot of pedagogical value, we outline the proof that the author had given in [22]. In that work, both the PS-action as well as the Polyakov-Liouville action of Eq. (22.51) have been analysed. We start by illustrating the case of S_0^{PS} .

The action S_0^{PS} is clearly invariant under $\delta_b X^\mu(\sigma, \tau) = b^\mu$ where b^μ does not depend on the world-sheet coordinates (σ, τ) . Therefore this is a case of a global invariance and Nöther's theorem can be applied straightforwardly. The trick to compute the corresponding conserved quantities is to consider a (σ, τ) -dependent $b^\mu(\sigma, \tau)$. The variation of the action, which is no longer zero, will then be of the form

$$\delta_b S_0^{PS} = \frac{1}{4\pi a^2} \int d\tau^+ d\tau^- \{ \partial_+ b \cdot \partial_- X + \partial_+ X \cdot \partial_- b \} \quad (22.72)$$

The conserved momentum densities are given by

$$\delta_b S_0^{PS} \equiv \int (p_+ \cdot \partial_- b + p_- \cdot \partial_+ b) \quad (22.73)$$

Therefore

$$p_+^\mu = \frac{1}{4\pi a^2} \partial_+ X^\mu \quad p_-^\mu = \frac{1}{4\pi a^2} \partial_- X^\mu \quad (22.74)$$

The total conserved momentum is obtained by integrating p_τ over all σ (see Chap. 17 for similar details in string theory). This is given by

$$p_\tau^\mu(\sigma, \tau) = p_+^\mu + p_-^\mu \quad (22.75)$$

On expanding X^μ around the classical background $X_{cl}^\mu = R(e_+^\mu \tau^+ + e_-^\mu \tau^-)$ i.e. $X^\mu = X_{cl}^\mu + Y^\mu$, and using the mode expansion for Y^μ owing to the EOM $\partial_{+-} Y^\mu = 0$

$$Y^\mu(\sigma, \tau) = q^\mu + (a\alpha_0^\mu \tau^- + ia \sum_{m \neq 0} \frac{\alpha_m^\mu}{m} e^{-im\tau^-}) + (a\tilde{\alpha}_0^\mu \tau^+ + ia \sum_{m \neq 0} \frac{\tilde{\alpha}_m^\mu}{m} e^{-im\tau^+}) \quad (22.76)$$

It is easy to show that

$$p_\tau^\mu(\sigma, \tau) = \frac{R}{4\pi a^2} (e_+^\mu + e_-^\mu) + \frac{1}{4\pi a} (\alpha_0^\mu + \tilde{\alpha}_0^\mu) + \dots \quad (22.77)$$

The ... above integrate to zero, giving the expression for the total momentum (for closed strings)

$$p^\mu = \frac{R}{2a^2}(e_+^\mu + e_-^\mu) + \frac{1}{2a}(\alpha_0^\mu + \tilde{\alpha}_0^\mu) \quad (22.78)$$

After this warm-up exercise with the much simpler S_0^{PS} , we now turn to the case of the PS-action. Following the same procedures as in the case of S_0^{PS} , the results for the PS-action are:

$$\begin{aligned} \Delta p_+^{ps\mu} = & \frac{\beta}{4\pi} \left\{ \frac{\partial_{++} X^\mu \partial_{--} X \cdot \partial_+ X}{L^2} - \frac{2 \partial_+ X^\mu \partial_{++} X \cdot \partial_- X \partial_{--} X \cdot \partial_+ X}{L^3} \right. \\ & \left. - \partial_- \left(\frac{\partial_+ X^\mu \partial_{++} X \cdot \partial_- X}{L^2} \right) \right\} \end{aligned} \quad (22.79)$$

with a similar expression for $\Delta p_-^{ps\mu}$. As we shall see, a separation of Y^μ into $Y_0^\mu + Y_1^\mu$ as before is not even necessary, nor any mode expansions. Instead, the general form of all functions of (τ^+, τ^-) is

$$Y^\mu(\tau^+, \tau^-) = F^\mu(\tau^+) + G^\mu(\tau^-) + H^\mu(\tau^+, \tau^-) \quad (22.80)$$

where H^μ is such that it has no purely *holomorphic* (i.e. functions of τ^- only) or *anti-holomorphic* (functions of τ^+ only) parts. From the EOM to order R^{-3} it follows that (this can also be obtained from Eq. (22.69) on substituting $Y_0 = F + G$ and $Y_1 = H$ and integrating it once more wrt to τ^-):

$$H^\mu = \frac{2\beta a^2}{R^3} [-e_- \cdot \partial_{++} F \partial_- G^\mu - e_+ \cdot \partial_{--} G \partial_+ F^\mu + e_+^\mu \partial_+ F \cdot \partial_{--} G + e_-^\mu \partial_{++} F \cdot \partial_- G] \quad (22.81)$$

Using these, the corrections to the momentum densities arising from the PS-action can be shown to be (with the short-hand notation $F_{++} = \partial_{++} F$ etc.)

$$\begin{aligned} \Delta p_+^{ps\mu} &= \frac{\beta}{2\pi R^3} \{ e_-^\mu F_{+++} \cdot G_- - e_+^\mu F_{++} \cdot G_{--} - G_-^\mu e_- \cdot F_{+++} + F_{++}^\mu e_+ \cdot G_{--} \} \\ \Delta p_-^{ps\mu} &= \frac{\beta}{2\pi R^3} \{ e_+^\mu G_{---} \cdot F_+ - e_-^\mu G_{--} \cdot F_{++} - F_{++}^\mu e_+ \cdot G_{---} + G_{--}^\mu e_- \cdot F_{++} \} \end{aligned} \quad (22.82)$$

Rather remarkably, these equations can be re-expressed as

$$\Delta p_+^{ps\mu} = \frac{\beta}{2\pi R^3} \partial_+ C^\mu \quad \Delta p_-^{ps\mu} = -\frac{\beta}{2\pi R^3} \partial_- C^\mu \quad (22.83)$$

with

$$C^\mu = e_-^\mu F_{++} \cdot G_- - e_+^\mu F_+ \cdot G_{--} + e_+ \cdot G_{--} F_+^\mu - e_+ \cdot F_{++} G_{--}^\mu \quad (22.84)$$

In other words, the corrections of order R^{-3} to the momentum densities of the PS-action are of the so called *improvement* type, whose contribution to the total momentum vanishes identically! This is so because $\Delta p_\tau^{ps\mu}$ has the form of a derivative wrt σ and Y^μ vanishes at the boundaries of the σ -integration. Thus the total momentum does not receive any corrections of order R^{-3} i.e. they are $O(R^{-4})$.

22.5 Covariant Calculus for Effective String Theories

As already mentioned, the Polyakov-Liouville action

$$S_{PL} = \frac{26-D}{48\pi} \int d\tau^+ d\tau^- \frac{\partial_+ L \partial_- L}{L^2} \quad (22.85)$$

is *exactly* invariant under (induced by the infinitesimal transformations $\tau^\pm \rightarrow \tau^\pm - \epsilon^\pm(\tau^\pm)$)

$$\delta_+ X^\mu = \epsilon^+(\tau^+) \partial_+ X^\mu \quad \delta_- X^\mu = \epsilon^-(\tau^-) \partial_- X^\mu \quad (22.86)$$

This should be contrasted with only the approximate invariance of Eq. (22.52) under the approximate transformations of Eq. (22.53). We have already remarked how the algebra of Eq. (22.53) closes only approximately while the algebra of Eq. (22.36) closes exactly. If one had followed the strategy of PS to higher orders, one would have to continually adjust the actions and transformation laws. On the contrary, our strategy, detailed in [20,21], is to keep the transformation laws *fixed* to be that of Eq. (22.36), and find a systematic calculus to generate actions invariant under it.

It is worth emphasizing the actual nature of the *exact* invariances of actions like S_{PL} and even the simpler S_0^{PS} . Denoting their integrands (the Lagrangean densities) by $\mathcal{L}_0^{PS}$ and $\mathcal{L}_{PL}$ it is easy to check that under, say $\epsilon^+(\tau^+)$ transformations

$$\delta_+ \mathcal{L}_0^{PS} = \partial_+(\epsilon^+ \mathcal{L}_0^{PS}) \quad \delta_+ \mathcal{L}_{PL} = \partial_+(\epsilon^+ \mathcal{L}_{PL}) + \partial_-(\partial_+^2 \epsilon^+ \mathcal{L}_{PL}) \quad (22.87)$$

In other words, only the respective actions are invariant. These are typical of general coordinate variations.

Another important aspect which has already been discussed is that the PS and PL actions are related by a *field redefinition*. This guarantees that the physical predictions of the two are the same (upto some order in R^{-1}), though the transformation laws take very different forms. If in our new strategy the transformation laws are going to be fixed, no field redefinitions will be allowed any more. This also means that the various means of simplifying the actions that were initially proposed by PS, and subsequently employed by both Drummond and us, can no longer be used. However, it may often prove useful to resort to field redefinitions after the relevant conserved quantities have been obtained. Needless to say, great care must be exercised to maintain consistency.

So far, our remarks have only been with regard to the PS-type effective theories whose leading order action S_0 was obtained from the Nambu-Goto action S_{NG} by fixing the so called conformal gauge of Eq. (22.31). The Nambu-Goto action enjoyed the much larger *world-sheet general coordinate invariance* while after fixing the conformal gauge, the *residual* general coordinate transformations that leave the conformal-gauge action invariant are precisely those of Eq. (22.36). The Nambu-Goto action could serve as the starting point for both the Lüscher-Weisz effective theory (in the static gauge) and the PS-effective action in the conformal gauge.

This suggests to base the starting point for the covariant calculus on the Nambu-Goto type actions invariant under the full world-sheet general coordinate invariance

although the description does not involve any *intrinsic metric* on the world sheet. Another starting point would be the Polyakov description with an intrinsic metric for the world-sheet included, but whose invariances are extended to both world-sheet general coordinate invariance as well as *local* Weyl invariance, as already introduced in Eqs. (22.48) and (22.49). We shall separately discuss both. We shall essentially follow [20,21] after correcting some errors (which we shall point out in what follows) and presenting the results in a different, logically more transparent way.

22.5.1 Covariant Calculus I: The Nambu-Goto way

The crux of this approach (see [20,21] for more details) is that general coordinate invariance can be implemented even in the absence of any intrinsic world-sheet metric as long as objects can be constructed that have the same transformation property under general coordinate transformations as an intrinsic metric. The *induced metric* (already introduced in Eq. (22.30)) is indeed one such:

$$g_{\alpha\beta} = \partial_\alpha X^\mu \partial_\beta X_\mu \quad (22.88)$$

The reader is referred to [20,21] for other important aspects of this choice. In what follows, both for the Nambu-Goto way and the Polyakov way, invariance under *Target Space* Poincare transformations of the X^μ are always assumed. It should however be remembered that static gauges break this invariance manifestly. The infinitesimal forms of the world-sheet general coordinate transformations $\xi^\alpha \rightarrow \xi^\alpha - \epsilon^\alpha(\xi)$ induce the transformations

$$\delta_{gen}(\epsilon) X^\mu = \epsilon^\alpha \partial_\alpha X^\mu \quad (22.89)$$

satisfying the algebra

$$[\delta_{gen}(\epsilon_1), \delta_{gen}(\epsilon_2)] = (\epsilon_1^\beta \partial_\beta \epsilon_2^\alpha - \epsilon_2^\beta \partial_\beta \epsilon_1^\alpha) \partial_\alpha X^\mu \quad (22.90)$$

It is easy to see that the transformations of Eq. (22.36) are special cases of these.

Now the construction of generally covariant actions essentially follows their analogs in General Relativity. This involves the construction of the Riemann-Christoffel symbols (see Eq. (4.6.8) of Weinberg's book on Gravitation and Cosmology [23])

$$\Gamma_{\alpha\beta}^\delta = \frac{g^{\delta\kappa}}{2} \{ \partial_\alpha g_{\kappa\beta} + \partial_\beta g_{\kappa\alpha} - \partial_\kappa g_{\alpha\beta} \} \quad (22.91)$$

and *Covariant Derivatives* (see Eqs. 4.6.8 and 4.6.10 of [23])

$$V_{\beta;\alpha} \equiv D_\alpha V_\beta = \partial_\alpha V_\beta - \Gamma_{\alpha\beta}^\delta V_\delta; \quad V^\beta{}_{;\alpha} \equiv D_\alpha V^\beta = \partial_\alpha V^\beta + \Gamma_{\alpha\delta}^\beta V^\delta \quad (22.92)$$

Generalizations to arbitrary tensors can be found in [23]. Covariant derivatives involving only the metric tensor and the Christoffel connection are the Riemann-Curvature tensors(see [23]). In two dimensions, these take the particularly simple form

$$R_{\alpha\beta\gamma\delta} = \frac{R}{2}(g_{\alpha\gamma}g_{\beta\delta} - g_{\alpha\delta}g_{\beta\gamma}) \quad (22.93)$$

with R being the curvature scalar. Thus one need to consider R and it's covariant derivatives only.

An important relation, called the metric condition

$$g_{\alpha\beta;\gamma} = 0 \quad (22.94)$$

follows automatically from the various formulae given above.

Covariant actions can be built out of the scalar curvature and other curvature invariants though the simplest choice

$$I_{GB} = \sqrt{-g} R \quad (22.95)$$

leads to action that is a topological invariant in two dimensions. A particularly intriguing covariant action involving only R is the Polyakov-Liouville action itself, written in a manifestly covariant form

$$S_{Polya} = \int d^2\xi (\sqrt{-g} R) \left\{ \frac{1}{\nabla^2} (\sqrt{-g} R) \right\}(\xi) \quad (22.96)$$

where ∇^2 is the scalar-Laplacian. We shall comment on this later.

The covariant actions that can be built manifestly out of covariant derivatives of X^μ take the generic form

$$I_{cov2} = \sqrt{-g} D_{\alpha_1\beta_1\dots} X^{\mu_1} D_{\alpha_2\beta_2\dots} X^{\mu_2} \dots D_{\alpha_n\beta_n\dots} X^{\mu_n} A^{\alpha_1\beta_1\dots\alpha_2\beta_2\dots} B_{\mu_1\mu_2\dots} \quad (22.97)$$

with A composed of two-dimensional Levi-Civita symbols as well as the metric $g_{\alpha\beta}$ and B composed of the invariant tensors of the target space. Drummond [17] had proposed the following four actions arising at order R^{-6} and higher (recall that $L = \partial_+ X \cdot \partial_- X$):

$$\begin{aligned} M_1 &= \frac{1}{L^3} \partial_+^2 X \cdot \partial_+^2 X \partial_-^2 X \cdot \partial_-^2 X \\ M_2 &= \frac{1}{L^3} \partial_+^2 X \cdot \partial_-^2 X \partial_+^2 X \cdot \partial_-^2 X \\ M_3 &= \frac{1}{L^4} \partial_+^2 X \cdot \partial_-^2 X \partial_- X \cdot \partial_+^2 X \partial_+ X \cdot \partial_-^2 X \\ M_4 &= \frac{1}{L^5} (\partial_- X \cdot \partial_+^2 X)^2 (\partial_+ X \cdot \partial_-^2 X)^2 \end{aligned} \quad (22.98)$$

We illustrate the possible covariantisation of the first two of these:

$$\begin{aligned}\mathcal{M}_1 &= \sqrt{g} D_{\alpha_1\beta_1} X \cdot D_{\alpha_2\beta_2} X D^{\alpha_1\beta_1} X \cdot D^{\alpha_2\beta_2} X \\ \mathcal{M}_2 &= \sqrt{g} D_{\alpha_1\beta_1} X \cdot D^{\alpha_1\beta_1} X D_{\alpha_2\beta_2} X \cdot D^{\alpha_2\beta_2} X\end{aligned}\quad (22.99)$$

We discuss the systematic generation of higher order actions shortly.

Now these covariant actions can be worked out in various gauges of interest like the static gauge, conformal gauge etc. We shall return to those issues after we have developed the covariant formalism for the Polyakov approach with intrinsic metric.

22.5.2 Covariant Calculus II: The Polyakov Way

In the leading order of this approach the action is given by Eq. (22.48) which is invariant under the local Weyl-scalings of Eq. (22.49) over and above the invariance under world-sheet reparametrizations (general coordinate invariance). Now we show how to extend Polyakov's considerations for effective string theories involving higher derivatives of both X^μ as well as the intrinsic metric $h_{\alpha\beta}$. All the details can be found in [20,21].

It turns out that this extension requires considerably more powerful technical and conceptual tools than what Polyakov had to do. Since higher derivatives of the intrinsic metric are involved, it becomes necessary to introduce covariant derivatives for local Weyl-scalings also. In effect this amounts to new *connections* which play the role of *gauge fields* for local Weyl-invariance.

One starts by introducing the notion of a *Weyl-scaling dimension*, also called Weyl-weight, for every field. A world-sheet tensor ϕ of Weyl-scaling dimension j transforms under local Weyl-scaling according to

$$\phi(\xi) \rightarrow \phi'(\xi) = \omega(\xi)^j \phi(\xi) \quad (22.100)$$

Accordingly the intrinsic metric $h_{\alpha\beta}$ has Weyl-scaling dimension of 1. Now consider, for example, a world-sheet vector V_β with Weyl-weight j_V . The covariant derivative of V_β with respect to the reparametrisations is now given in terms of the $h_{\alpha\beta}$ unlike the Eq. (22.92):

$$V_{\beta;\alpha} \equiv D_\alpha V_\beta = \partial_\alpha V_\beta - \Gamma_{\alpha\beta}^\delta V_\delta \quad (22.101)$$

with the Christoffel connection now given in terms of the intrinsic metric:

$$\Gamma_{\alpha\beta}^\delta = \frac{h^{\delta\kappa}}{2} \{ \partial_\alpha h_{\kappa\beta} + \partial_\beta h_{\kappa\alpha} - \partial_\kappa h_{\alpha\beta} \} \quad (22.102)$$

Now the covariant derivative, as it stands, does not have a well-defined Weyl-weight. There are two sources to this difficulty; one is that derivatives of V_β are involved, and the other is that derivatives of $h_{\alpha\beta}$ are also involved through $\Gamma_{\alpha\beta}^\delta$.

This suggests the introduction of a *Weyl-covariant derivative* of tensors that will have the *same* Weyl-weight as the tensors themselves. From the experiences with gauge theories this is chosen to be of the form

$$\Delta_\alpha \phi \equiv \partial_\alpha \phi - j \chi_\alpha \phi \quad (22.103)$$

where χ_α transforms like a world-sheet vector. The mathematical requirement of Weyl-covariance of Δ_α can be stated as:

$$(\Delta_\alpha \phi)' = \omega^j \Delta_\alpha \phi \quad (22.104)$$

It is not difficult to see that this can be achieved provided χ_α transforms under Weyl-scalings as

$$\chi'_\alpha = \chi^\alpha + \partial_\alpha \ln \omega \quad (22.105)$$

which is indeed the way an abelian gauge-field transforms. This at once leads to a natural generalization of the Christoffel connection of Eq. (22.102) to something more appriate for the present context:

$$G^\delta_{\alpha\beta} = \frac{h^{\delta\kappa}}{2} (\Delta_\alpha h_{\kappa\beta} + \Delta_\beta h_{\kappa\alpha} - \Delta_\kappa h_{\alpha\beta}) \quad (22.106)$$

It will turn out to be convenient to split G according to

$$G^\delta_{\alpha\beta} = \Gamma^\delta_{\alpha\beta} + W^\delta_{\alpha\beta} \quad W^\delta_{\alpha\beta} = \frac{1}{2} (h_{\alpha\beta} \chi^\delta - \delta^\delta_\alpha \chi_\beta - \delta^\delta_\beta \chi_\alpha) \quad (22.107)$$

Since $\Gamma^\delta_{\alpha\beta}$ transforms like a *connection*, and $W^\delta_{\alpha\beta}$ transforms like a *tensor*, this split helps in establishing the important property that $G^\delta_{\alpha\beta}$ in itself transforms like a connection. From it's very construction, it is easy to see that the Weyl-weight of $G^\delta_{\alpha\beta}$ is actually zero (in other words, it is invariant under local Weyl-scalings) though neither $\Gamma^\delta_{\alpha\beta}$ nor $W^\delta_{\alpha\beta}$ has any well-defined Weyl-weight!

With all these preliminaries, we now give the construction of what we called *Weyl-reparametrisation covariant derivatives* of, say, a world-sheet tensor $T_{\beta_1 \dots \beta_n}$ of Weyl-weight j (generalization to arbitrary tensor types follows exactly as in General Relativity (GR)):

$$\mathcal{D}_\alpha T_{\beta_1 \dots \beta_n} \equiv \Delta_\alpha T_{\beta_1 \dots \beta_n} - G^\kappa_{\alpha\beta_1} T_{\kappa\beta_2 \dots \beta_n} - \dots - G^\kappa_{\alpha\beta_n} T_{\beta_1 \dots \beta_{n-1}\kappa} \quad (22.108)$$

Another representation of the same object that will be useful is

$$\mathcal{D}_\alpha T_{\beta_1 \dots \beta_n} = D_\alpha T_{\beta_1 \dots \beta_n} - j \chi_\alpha T_{\beta_1 \dots \beta_n} - W^\kappa_{\alpha\beta_1} T_{\kappa\beta_2 \dots \beta_n} - \dots - W^\kappa_{\alpha\beta_n} T_{\beta_1 \dots \beta_{n-1}\kappa} \quad (22.109)$$

In the first representation, every term has the same Weyl-weight as that of $T_{\beta_1 \dots \beta_n}$ while none of them transforms like a tensor under reparametrisations. In the second representation, every term transforms as a tensor under reparametrisations while

none of them has a definite Weyl-weight! Putting together, one concludes that $\mathcal{D}T$ transforms covariantly under both reparametrisations as well as Weyl-scalings.

With the help of these Weyl-reparametrisation covariant derivatives one can proceed to construct actions that have Weyl-weights zero and that also transform as scalar densities under reparametrisations. We gather below some important properties of the variety of covariant derivatives constructed so far. They all obey the Leibnitz rule:

$$D_\alpha T_1 T_2 = D_\alpha T_1 \cdot T_2 + T_1 \cdot D_\alpha T_2 \quad (22.110)$$

with identical relations for Δ_α and $\mathcal{D}_\alpha$. $\mathcal{D}$ also satisfies the metric condition

$$\mathcal{D}_\alpha h_{\beta\gamma} = 0 \quad (22.111)$$

Once again, there are two classes of manifestly covariant actions that can be constructed. The first of these is the exact analogs of the actions built with the Riemann curvature tensor in the Covariant Calculus-I. The generalized curvature tensor can be constructed in complete parallel (as for example by evaluating $[\mathcal{D}_\alpha, \mathcal{D}_\beta]$ on generic tensor fields). The explicit form is given by:

$$\mathcal{R}^\alpha_{\beta\gamma\delta} = \Delta_\gamma G^\alpha_{\beta\gamma} - \Delta_\delta G^\alpha_{\beta\gamma} + G^\alpha_{\gamma\eta} G^\eta_{\delta\beta} - G^\alpha_{\delta\eta} G^\eta_{\gamma\beta} \quad (22.112)$$

The other is through various Weyl-reparametrisation covariant derivatives of X^μ .

22.5.3 Weyl Connections and Weyl-Weight Compensators

We have so far not specified what we called the *Weyl Connection*, χ_α , so far. As long as it transforms as a world-sheet vector and satisfies Eq. (22.105), its specific form was not required. The main objective of our covariant calculus is to produce, in a systematic manner, actions that are scalar densities under reparametrisations, and are Weyl-scaling(local) invariant i.e. they should carry zero Weyl-weights. While the former can be achieved with the help of what has been developed so far in this chapter, the latter still needs further scrutiny.

In fact any χ_α of the form

$$\chi_\alpha = \frac{\partial_\alpha \ln \Phi}{W_\Phi} \quad (22.113)$$

with Φ a world-sheet scalar of Weyl-weight W_Φ would be a good candidate. Clearly, this still leaves many possibilities open for (Φ, W_Φ) . To appreciate that, let us recast the above equation in a more suggestive way as $\mathcal{D}_\alpha \Phi = 0$. That there are non-trivial solutions to this follows on noting that $\mathcal{L} = h^{\alpha\beta} \partial_\alpha X \cdot \partial_\beta X$ indeed satisfies this with $W_{\mathcal{L}} = -1$. By the Leibnitz rules discussed earlier all choices of the type $\Phi_n = \mathcal{L}^n$ with $W_{\Phi_n} = -n$ are equally good candidates. In fact, given any two candidates (Φ_1, W_{Φ_1}) and (Φ_2, W_{Φ_2}) , it is easy to see that $(\Phi_1 \Phi_2, W_{\Phi_1} + W_{\Phi_2})$ is also a candidate!

To highlight the many issues involved, let us consider a contravariant vector V^α with Weyl-weight J (since covariant and contravariant tensors are mapped using $h_{\alpha\beta}$ and its inverse, their Weyl-weights are in general different), and its Weyl-reparametrisation covariant derivative

$$\mathcal{D}_\alpha V^\beta = \nabla_\alpha V^\beta - J\chi_\alpha V^\beta + W_{\alpha\gamma}^\beta V^\gamma \quad (22.114)$$

Hence

$$\mathcal{D}_\alpha V^\alpha = \nabla_\alpha V^\alpha - J\chi_\alpha V^\alpha + W_{\alpha\gamma}^\alpha V^\gamma \quad (22.115)$$

The reparametrisation part is given by $\nabla_\alpha V^\alpha = \frac{1}{\sqrt{h}} \partial_\alpha (\sqrt{h} V^\alpha)$. In fact this is what gives the result that $\sqrt{h} \nabla_\alpha V^\alpha$ is actually an ordinary four-divergence, a very important result in proving charge conservation in GR. We wish to establish a similar result for $\sqrt{h} \mathcal{D}_\alpha V^\alpha$. On using Eq. (22.107), it follows that $W_{\alpha\gamma}^\gamma = -\chi_\alpha$, and, on using Eq. (22.113), one obtains

$$\mathcal{D}_\alpha V^\alpha = \frac{\Phi^{\frac{J+1}{W_\Phi}}}{\sqrt{h}} \partial_\alpha (\sqrt{h} \Phi^{-\frac{J+1}{W_\Phi}} V^\alpha) \quad (22.116)$$

This expression brings out the central issue that scalar densities do not always come with zero Weyl-weights! Thus, $\sqrt{h} \mathcal{D}_\alpha V^\alpha$, though a scalar density under reparametrisations, has the non-zero Weyl-weight $J+1$. Now we see that if $\sqrt{h} \mathcal{D}_\alpha V^\alpha$ were also to maintain the ordinary four-divergence property, it has to be scaled (multiplied) by the exact factor of $\Phi^{-\frac{J+1}{W_\Phi}}$.

The possibility of $\Phi = \mathcal{L}$, $W_\Phi = -1$ mentioned earlier is the simplest choice that one can make. It also fulfills what we called a *Denominator Principle* in [20, 21]. It is clear that many actions for effective string theories will simply be unacceptable for a fundamental theory of strings as these actions will become singular for certain string configurations. But in the effective description actions are permissible as long as they are not singular for the classical string configuration X_{cl} in the above and small fluctuations around it. In the PS-effective action the denominator was L^2 which clearly satisfies the denominator principle. In the Drummond higher-order actions too all denominators were powers of L . $\mathcal{L}$ is the closest analog of L . We shall clarify this more as we go along.

Returning to the four-divergence (ordinary) from $\sqrt{h} \mathcal{D}_\alpha V^\beta$, with $\Phi = \mathcal{L}$ as the compensator, it is $\mathcal{L}^{J+1} \sqrt{h} \mathcal{D}_\alpha V^\alpha$ that is an ordinary four-divergence.

To throw more light on these issues, let us construct a manifestly covariant action under Covariant Calculus-II (the Polyakov way) that is analogous to, say, $\mathcal{M}_2$ of Eq. (22.99). A first guess would be

$$\tilde{\mathcal{N}}_1 = \sqrt{h} \{ \mathcal{D}_{\alpha_1\beta_1} X \cdot \mathcal{D}_{\alpha_2\beta_2} X h^{\alpha_1\alpha_2} h^{\beta_1\beta_2} \}^2 \quad (22.117)$$

Though a scalar-density under reparametrisations, it's Weyl-weight is -3, making it unsuitable for a covariant action. A Weyl-weight compensator of $\mathcal{L}^{-3}$ renders it into an acceptable action:

$$\mathcal{N}_1 = \sqrt{h} \mathcal{L}^{-3} \{ \mathcal{D}_{\alpha_1 \beta_1} X \cdot \mathcal{D}_{\alpha_2 \beta_2} X h^{\alpha_1 \alpha_2} h^{\beta_1 \beta_2} \}^2 \quad (22.118)$$

We shall see the consistency of the choice $\Phi = \mathcal{L}$ with the denominator principle when we examine the manifestly covariant actions in the light cone gauge.

22.6 Gauge Fixing the Covariant Actions

22.6.1 The Static Gauge

We make some brief remarks about the static gauge for the covariant actions. As is well known, quantisation requires gauges to be fixed. At a classical level, the effective actions can be worked out in this gauge exactly as was done for the Nambu-Goto action. Even this lead to an infinite sequence of actions the first few of which were analysed in the pioneering works of Lüscher and Weisz [1,3]. Even at the classical level we found how the underlying world-sheet reparametrisation invariance of the Nambu-Goto action led to very important constraints between the coefficients of the Lüscher-Weisz type of effective actions with only the transverse fields.

The same procedure can now be followed for the actions given by the Covariant Calculus-I. Though the calculations will become increasingly unwieldy as is typical of most higher order calculations, the procedure is unambiguous. It was tacitly assumed that the resulting effective actions are of a *local* type in a QFT-sense. But the manifestly covariant action of Eq. (22.96) raises some reservations in this respect. In the conformal gauge, this led to the local action of Eq. (22.51) which was at the heart of all the important conclusions reached by the PS-effective string approach. The chief among them were the validity of the truncated Arvis potential for $D \neq 26$, and, the absence of any corrections at the level of R^{-3} to this potential. Consequently, this action must play an equally important role in the static gauge calculations too. But it is hard to see how this can produce local actions at all orders. The extensive works by Aharony and collaborators (more on them later) in this gauge should clarify this vexing issue.

22.6.2 Covariant Calculus I: The Conformal Gauge

The conformal gauge in this case, as already described before, is defined by $g_{++} = g_{--} = 0$, and, $g_{+-} = g_{-+} = \partial_+ X \cdot \partial_- X = L$. The inverse of the metric is given by $g^{+-} = g^{-+} = L^{-1}$. The denominator principle mentioned before can also be taken as the requirement that the metric be non-singular.

The non-vanishing components of the Christoffel connection and the curvature tensor are easily worked out:

$$\begin{aligned}\Gamma_{++}^{(1)+} &= \partial_+ \ln L & \Gamma_{--}^{(1)-} &= \partial_- \ln L \\ R_{+-+}^+ &= R_{-+-}^- = -R_{++-}^+ = -R_{--+}^- = \partial_{+-} \ln L\end{aligned}\quad (22.119)$$

where we have used the superscript (1) to denote that the Christoffel symbols pertain to Covariant Calculus I and are calculated with the induced metric. The resulting scalar curvature is

$$R = -2 \frac{\partial_{+-} \ln L}{L} \quad \sqrt{g} R = -2 \partial_{+-} \ln L \quad (22.120)$$

Polchinski and Strominger, in arriving at their effective action, started with the Liouville action and substituted for the conformal factor e^ϕ the component g_{+-} of the induced metric. This certainly appeared ad hoc. Our covariant calculus avoids making such jumps. By substituting Eq. (22.120) in the non-local effective action where by R one means the scalar curvature evaluated from $g_{\alpha\beta}$, and on noting that the scalar Laplacian in this gauge is just ∂_{+-} , one straight away recovers S_{PL} .

We end this subsection by writing down the explicit expressions for a few non-trivial covariant derivatives of X^μ . Others can be easily worked out using the expressions for the Christoffel connection.

$$\begin{aligned}D_{++}X^\mu &= \partial_{++}X^\mu - \partial_+ \ln L \partial_+ X^\mu & D_{--}X^\mu &= \partial_{--}X^\mu - \partial_- \ln L \partial_- X^\mu \\ D_{+-}X^\mu &= D_{-+}X^\mu = \partial_{+-}X^\mu\end{aligned}\quad (22.121)$$

22.6.3 Covariant Calculus II: The Conformal Gauge

Now we turn to a discussion of the conformal gauge for the Covariant Calculus II. The details are fascinatingly different from those for Covariant Calculus I! The local Weyl-rescaling allows the entire $h_{\alpha\beta}$ to be set equal to the *flat* metric! We give the details of this conceptually very important aspect. The infinitesimal reparametrisations now act on both (X, h) according to

$$\delta_\epsilon X^\mu = \epsilon^\alpha \partial_\alpha X^\mu; \quad \delta_\epsilon h^{\alpha\beta} = \epsilon^\gamma \partial_\gamma h^{\alpha\beta} - \partial_\gamma \epsilon^\alpha h^{\gamma\beta} - \partial_\gamma \epsilon^\beta h^{\alpha\gamma} \quad (22.122)$$

The infinitesimal Weyl-scalings do not change X^μ but act on $h_{\alpha\beta}$ according to

$$\delta_\lambda X^\mu = 0 \quad \delta_\lambda h_{\alpha\beta} = \delta \lambda h_{\alpha\beta} \quad (22.123)$$

It is known from GR that in two dimensions, every intrinsic metric can be brought to a *conformally flat* metric. The local Weyl-rescaling can then be used to bring the intrinsic metric to a *flat* metric. It is important to appreciate that both these require *large*, and not infinitesimal, transformations. Thus in the conformal gauge for Calculus-II, first one brings the intrinsic metric to $h_{\alpha\beta} = \eta_{\alpha\beta}$.

Introducing the coordinates $\tau^\pm = \tau \pm \sigma$, the non-vanishing components of the fully fixed intrinsic metric becomes

$$h_{+-} = h_{-+} = 2 \quad h^{+-} = h^{-+} = \frac{1}{2} \quad (22.124)$$

As can easily be verified, this does not fix the freedom to make reparametrisations and local Weyl-scalings fully. Combined infinitesimal reparametrisations ϵ_α and infinitesimal Weyl-scalings $\delta \lambda$ such that

$$\delta \lambda \eta_{\alpha\beta} = \partial_\alpha \epsilon_\beta + \partial_\beta \epsilon_\alpha \quad (22.125)$$

still preserve $h_{\alpha\beta} = \eta_{\alpha\beta}$. These *residual* transformations constitute the conformal transformations in Calculus-II.

In this conformal gauge, with the coordinates $\tau^\pm$, $\mathcal{L} = \partial_+ X \cdot \partial_- X = L$. Consequently,

$$\chi_+ = -\partial_+ \ln L \quad \chi_- = -\partial_- \ln L \quad (22.126)$$

Furthermore, the Christoffel connections of Eq. (22.102) now vanish identically (as they should, for flat space!):

$$\Gamma^{(2)\gamma}_{\alpha\beta} = 0 \quad (22.127)$$

where we have used the superscript (2) to distinguish these Christoffel connections from those of Calculus-I. The $W^\gamma_{\alpha\beta}$ in this conformal gauge are given by (there is a typo in [20,21]):

$$W^\gamma_{\alpha\beta} = -\frac{1}{2L} (\eta_{\alpha\beta} \partial^\gamma - \delta^\gamma_\alpha \partial_\beta - \delta^\gamma_\beta \partial_\alpha) L \quad (22.128)$$

It's components explicitly written down are:

$$W^+_{++} = -\chi_+ = \partial_+ \ln L \quad W^-_{--} = -\chi_- = \partial_- \ln L \quad (22.129)$$

It is instructive to examine the structure of $\mathcal{D}_\alpha$ in this gauge. Consider a tensor $T_{\alpha\beta\dots}$ of Weyl-weight j . The components of it's Weyl-reparametrisation covariant derivatives are:

$$\mathcal{D}_+ T^{(j)}_{\dots} = \partial_+ T^{(j)}_{\dots} - j \chi_+ T^{(j)}_{\dots} - t_+ W^+_{++} T^{(j)}_{\dots} \quad (22.130)$$

where t_+ is the number of + indices of T . Substituting the value of W^+_{++} from above, this takes the rather simple form

$$\mathcal{D}_+ T^{(j)}_{\dots} = \partial_+ T^{(j)}_{\dots} + (t_+ - j) \chi_+ T^{(j)}_{\dots} \quad (22.131)$$

and a similar expression for $+$ $\leftrightarrow$ $-$.

Now one can work out the expressions for various covariant derivatives of X^μ and construct the required actions in this covariant gauge for Calculus-II, but we will prove a powerful result in the next section demonstrating the complete equivalence of the conformal gauges for both classes of the covariant calculus.

22.7 Equivalence of Conformal Gauges

We shall now demonstrate that the Weyl-reparametrisation covariant derivatives $\mathcal{D}_{\alpha\beta\dots}X^\mu$ are the same in value as the covariant derivatives $D_{\alpha\beta\dots}X^\mu$. The equivalence is first shown for the larger class of tensors with Weyl-weight zero, of which X^μ are the simplest examples. Let $T_{\beta_1\dots\beta_n}$ be such a zero Weyl-weight tensor. Then,

$$\begin{aligned}\mathcal{D}_\alpha T_{\beta_1\dots\beta_n} &= \partial_\alpha T_{\beta_1\dots\beta_n} - G_{\alpha\beta_1}^\gamma T_{\gamma\beta_2\dots} - \dots \\ &= \partial_\alpha T_{\beta_1\dots\beta_n} - W_{\alpha\beta_1}^\gamma T_{\gamma\beta_2\dots} - \dots \\ &= \partial_\alpha T_{\beta_1\dots\beta_n} - \Gamma_{\alpha\beta_1}^{(1)\gamma} T_{\gamma\beta_2\dots} - \dots \\ &= D_\alpha T_{\beta_1\dots\beta_n}\end{aligned}\tag{22.132}$$

The first step was just invoking the definition of $\mathcal{D}$; the second step made use of the fact that all components of $\Gamma^{(2)}$ are zero; the third step made use of the fact that all the components of $W_{\alpha\beta}^\gamma$ are equal to those of $\Gamma_{\alpha\beta}^{(1)\gamma}$.

This is a striking result and has many important consequences. One of these is that all the components of $\mathcal{R}_{\beta\gamma\delta}^\alpha$ equal those of $R_{\beta\gamma\delta}^\alpha$. It is worth clarifying this equality in some more detail. It may superficially appear that $\mathcal{R}$ defined, say, through

$$[\mathcal{D}_\alpha, \mathcal{D}_\beta] V_\mu^{(j)} = -\mathcal{R}_{\mu\beta\alpha}^{(j)\sigma} V_\sigma^{(j)}\tag{22.133}$$

may have a j -dependence. On writing $V_\mu^{(j)} = \Phi^{\frac{j}{W_\Phi}} V_\mu$ where V_μ is a zero Weyl-weight tensor, and on using $\mathcal{D}_\alpha \Phi = 0$ as noted before, it can easily be shown that $\mathcal{R}$ is j -independent. Furthermore, $\mathcal{R}$ is also a zero Weyl-weight tensor (these two statements have different meanings). Consequently, all the $\mathcal{D}$ covariant derivatives of $\mathcal{R}$ equal the covariant derivatives of the curvature tensor in Calculus-I. Therefore all actions constructed in the two different calculi are the same.

This is not just an accident, and can in fact be understood in a more general way. Consider another internal metric $h'_{+-} = \Phi h_{+-}$. This has the effect of making the new $\Phi' = 1$ and consequently $\chi'_\alpha = 0$. Therefore the transformed W vanishes i.e. $W'^\gamma_{\alpha\beta} = 0$. However the transformed Christoffel connection does not vanish anymore. Instead, $\Gamma^{(2)\gamma}_{\alpha\beta} = \Gamma^{(1)\gamma}_{\alpha\beta}$. However, the sum, $G = W + \Gamma^{(2)}$ is unchanged, which is not surprising as G is a tensor with zero Weyl-weight.

Now one can repeat the strategy we had adopted in proving the equivalence for $\mathcal{D}$ -derivatives of tensors with zero Weyl-weights. Even though we proved the equivalence for Weyl-reparametrisation covariant derivatives of tensors with zero Weyl-weights, the proof can be easily extended even to tensors with arbitrary Weyl-weights. This completes the proof of the exact equivalence of conformal gauges for the two Covariant Calculi. It must be emphasized that these equivalences hold at purely classical levels with the understanding that the calculus be used to construct various classical actions which are then to be quantized as was done in the earlier parts of this chapter.

As an explicit realization of this equivalence, consider Eq. (22.118). Evaluating it in the conformal gauge for Calculus-II yields:

$$\mathcal{N}_1 = \frac{1}{2L^3} (D_{++}X \cdot D_{--}X + D_{+-}X \cdot D_{+-}X)^2 \quad (22.134)$$

Likewise, evaluating $\mathcal{M}_2$ of Eq. (22.99) in the conformal gauge for Calculus-I gives:

$$\mathcal{M}_2 = \frac{4}{L^3} (D_{++}X \cdot D_{--}X + D_{+-}X \cdot D_{+-}X)^2 \quad (22.135)$$

Thus both represent the same action. The equality of all the covariant derivatives in the two calculi was crucial for this.

22.8 Drummond Actions as Examples

As a way of illustrating these ideas we show how covariantisation of the order R^{-6} actions proposed by Drummond [17, 19] works according to our Covariant Calculus-I. As clearly emphasized, in our covariant calculi the transformation laws, valid to all orders, are held fixed. Therefore, terms proportional to the leading order equations of motion (EOM) can not be thrown away as that would be tantamount to performing field redefinitions which would generically alter the transformation laws. However, Drummond had arrived at these candidate forms after dropping EOM terms and any comparison can only be made upto terms proportional to the EOM's.

The Covariant Calculus-I conformal gauge expressions for $\mathcal{M}_1$, $\mathcal{M}_2$ after ignoring terms proportional to $D_{+-}X$ (by Eq. (22.121)) these are proportional to leading order EOM) are:

$$\begin{aligned} \mathcal{M}_1 &= \frac{2}{L^3} \{D_{++}X \cdot D_{++}X D_{--}X \cdot D_{--}X + (D_{++}X \cdot D_{--}X)^2\} \\ \mathcal{M}_2 &= \frac{4}{L^3} (D_{++}X \cdot D_{--}X)^2 \end{aligned} \quad (22.136)$$

The combination $\mathcal{M}_1 - \frac{\mathcal{M}_2}{2}$ takes the form

$$\mathcal{M}_1 - \frac{\mathcal{M}_2}{2} = \frac{2}{L^3} (D_{++}X \cdot D_{++}X)(D_{--}X \cdot D_{--}X) \quad (22.137)$$

In accordance with the proposals made by PS for constructing higher-order actions, Drummond had also dropped the leading order constraints $\partial_{\pm} X \cdot \partial_{\pm} X$ and their derivatives. Modulo such terms the combination above is just the M_1 of Eq. (22.98). After some tedious algebra it can likewise be shown that

$$\mathcal{M}_2 = M_2 - 2M_3 + M_4 \quad (22.138)$$

More of tedious algebra shows that the covariant calculus can not produce any other combinations of M_i .

In summary, upto order R^{-6} the most general conformal gauge action possible is given by

$$S_6^{conf} = \int \frac{d^2\sigma}{4\pi} \left\{ \frac{L}{a^2} + \beta \frac{\partial_+ L \partial_- L}{L^2} + \beta_1 (M_2 - 2M_3 + M_4) + \beta_2 M_1 \right\} \quad (22.139)$$

It is useful to record the fully covariant version (which can be evaluated in any gauge) of this:

$$S_6^{cov} = \int \frac{d^2\sigma}{4\pi} \left\{ S_{NG} + \frac{\beta}{16\pi} (\sqrt{-g} R) \frac{1}{\nabla^2} (\sqrt{-g} R)(\xi) + \beta_2 \mathcal{M}_1 + (\beta_1 - \frac{\beta_2}{2}) \mathcal{M}_2 \right\} \quad (22.140)$$

This should be considered as one of the triumphs of our Covariant Calculus. This entire action is invariant, to all orders in R^{-1} under the simple transformations of Eq. (22.36). The first term *entirely* accounts for the Nambu-Goto action. This is unlike the case in the static gauge where just the Nambu-Goto action leads to an infinite number of action terms with increasing number of derivatives. The second term is the equivalent of the PS-action and as already elaborated fixes β at $\beta_c = \frac{D-26}{12}$. Consequently, the effect of this action is to produce *universal* terms to all orders, and in this case to R^{-5} order. Thus, for closed strings where one does not have to worry about any boundary term complications, the static potential and the spectrum to order R^{-6} (and most likely R^{-7}) is determined by only two parameters β_1, β_2 , which have no reason to be universal. These important conclusions follow without any detailed calculations. Detailed calculations in both the static gauge and the conformal gauge have upheld these expectations as will be discussed in the next section.

22.9 Spectrum of Effective Strings at Even Higher Orders

The natural question that arose after Drummond's, and, Hari Dass and Matlock's works that showed that the spectrum of effective strings upto order R^{-3} was of the same form as the truncated Arvis potential upto the same order, but now with the crucial difference being the validity in $D \neq 26$, was whether this remarkable feature would survive even higher order corrections. Here we separately discuss the results obtained by (i) Aharony and his collaborators on the one hand, and by (ii) the author and collaborators on the other. The narratives are not in chronological order.

22.9.1 Results by Aharony et al.

Aharony and Karzbrun were the first to announce results for the spectrum of effective string theories at order R^{-5} [24,25]. They chose to work in the static gauge that Lüscher and Weisz had used in their pioneering work in [1]. They essentially followed

LW by systematically generating higher-derivative action terms, assuming *locality*. Recall that LW had introduced two such terms with coefficients c_2, c_3 at the level of four-derivative terms, symbolically denoted as $\mathcal{L}_4$. Aharony and Karzbrun first write down the most general six-derivative terms which fall into two categories, $\mathcal{L}_{6,4}$ which involve four transverse fields (which they designate as X^i instead of h^i as done by LW), and, $\mathcal{L}_{6,6}$ with 6 X^i . The former is of two kinds with coefficients c_4, c_5 , of which they argue that c_5 terms do not contribute. The latter has three types of terms with coefficients c_6, c_7, c_8 of which they show that c_8 type of action is a linear combination of c_6 and c_7 types. They go on to show, rather like the situation with the c_2, c_3 terms that all but c_4 are contained in the Nambu-Goto action itself. They also study carefully constraints among the c_i in close parallel to the open-closed string duality of LW. Like LW, they also make extensive use of partition functions.

They too chose to work with closed strings avoiding possible complications arising from boundary terms. This is a highly detailed work examining strings in various backgrounds and is a treasure-house of techniques. An important fallout was that they too showed the *absence* of R^{-3} corrections to the spectrum of the Nambu-Goto action i.e. the Arvis spectrum in the static gauge also. Like LW, they too did not address the question of validity of the results for $D \neq 26$. More specifically, they did not carry out the consistency requirements as in string theory in the static gauge which is how the question of validity in $D \neq 26$ has to be addressed.

They could not fix the coefficient c_4 which was subsequently addressed by Aharony, Field and Klinghofer [26] who analysed the R^{-5} corrections within the PS-formalism that made use of many insights obtained by Drummond and us. We will come to that shortly. There is type of non-uniformity in Aharony and Karzbrun's treatment; while they use the fully covariant Nambu-Goto action to analyse part of the problem, they did not try to base the whole analysis on a systematic construction of fully covariant effective actions as done by us in our Covariant Calculus. As already mentioned, they would then have to face the problem of possible non-local actions in the static gauge. We shall return to this once more after commenting on the Polchinski-Strominger approach adopted by Aharony, Field and Klinghofer [26].

Now we comment on their work. They extend the R^{-3} analysis of Drummond and us to order R^{-5} within the PS-formalism, which they funnily call the *orthogonal* gauge while the correct nomenclature ought to be the *conformal gauge* or even the *covariant gauge*. This, however, does not dent the correctness of their analysis. Theirs is a more or less routine extension of the R^{-3} order analysis except for a few important technical differences. As already discussed, the correctness of the R^{-3} analysis hinged on the fact that the string momentum P^μ does not receive any corrections at this order. So these authors address the issue of higher order corrections to string momentum. By a clever choice of field redefinitions they show that P^μ does not get corrected at R^{-5} level also.

Then they construct the Virasoro generators to this order, which now involve quartic order expressions in the oscillators. To quantize, they adopt what they call a *Weyl Ordering*. They do not seem to have explicitly verified that they satisfy the correct Virasoro algebra. Their two most important results are: (i) that the ground state energy even to this order agrees with the Arvis potential expanded to this order

(two orders higher in R^{-1} than the truncated Arvis potential) but now valid for $D \neq 26$, and, ii) the spectrum of the (1, 1) excited states show a *deviation* from the Arvis spectrum to this order (except in $D = 3$).

Another important conclusion these authors reach is that their conformal gauge results to order R^{-5} are in full agreement with the earlier static gauge calculations of Aharony and Karzbrun provided their c_4 , the only parameter they had not fixed, is taken to be $c_4 = \frac{D-26}{192\pi} = \frac{\beta_c}{16\pi}$. This raises several interesting points and it clearly points to the correctness of both the R^{-5} order results. Firstly, it points to the origin of c_4 in the static gauge to be the same as that of the PS-action, equivalently what we called the Polyakov-Liouville action (actually Aharony, Field and Klinghofer based their analysis on this form rather than the form originally given by PS). This can only be consistent if the static gauge actions were also derived from the same parent fully covariant action i.e. of Eq. (22.96). But this covariant action was exactly *local* only in the conformal gauge. So one has to understand how it could have given rise to the local c_4 term in the static gauge. In particular, it is important to understand at what order, if at all, the static gauge starts producing non-local actions.

It is worth emphasizing that the results of Aharony, Field and Klinghofer are completely in accordance with the expectations of our Covariant Calculus-I as compactly encapsulated in Eqs. (22.139) and (22.140) [20, 21]. That is, at R^{-5} there are *no* free parameters and there is complete universality in the sense that the spectrum only depends on D . What is remarkable is that while the ground state energy even to this order only depends on D through the combination $(D - 2)$, reflecting the transverse nature of the physical degrees of freedom, the excited state energies do not. As per the Covariant Calculi, *potential* non-universal terms can only arise at order R^{-7} and higher, as also borne out by explicit calculations [25]. We say *potential* because these too could, in principle, show universality. As of now, the additional parameters β_1, β_2 of Eqs. (22.139) and (22.140) are free parameters.

22.9.2 Alleged Equivalence to Arvis Spectrum To All Orders

Around the same time as the work of Aharony and Karzbrun, this author, along with his collaborators Peter Matlock and Yashas Bharadwaj had made the rather extraordinary claims that (i) the spectrum to *all* orders of the Polyakov-Liouville action of Eq. (22.51) was *identical* to the Arvis Spectrum [27], (ii) that the inclusion of the two independent type of Drummond actions of Eq. (22.136) still does not change the spectrum from the Arvis spectrum [28], and, (iii) that inclusion of all the actions of Covariant Class-I also does not change the spectrum [29].

The author has himself been sceptical of these results but repeated and careful scrutiny has not revealed the fault lines. The author would like to highlight the main lines of reasoning so that this could be resolved. More so, because the methods used were very straightforward and mathematically elegant. It should be emphasized that the weakest link in the chain of reasoning is (i) itself as (ii) and (iii) depend on the correctness of (i).

So let us look at the salient features of (i). The first step consists in determining the equations of motion E^μ for the X^μ arising from this action, and in applying Nöther theorem to find the conserved T_{--} , to all orders, as a result of the exact invariance under Eq. (22.36) (likewise T_{++}). The reader is referred to [27] for all the formulae and details. As already noted in our analysis of the R^{-3} level corrections, this T_{--} is *superficially* not *holomorphic* i.e. a function of τ^- only as there are terms with $+$ -derivatives of fields. This issue is addressed by making use of Eq. (22.80) (recall that Y^μ , the fluctuation field, is defined through $X^\mu = X_{cl}^\mu + Y^\mu$) which we repeat here for ease of reading:

$$Y^\mu(\tau^+, \tau^-) = F^\mu(\tau^+) + G^\mu(\tau^-) + H^\mu(\tau^+, \tau^-) \quad (22.141)$$

On substituting this in T_{--} , one can decompose it into $T_{--} = T_{--}^h + T_{--}^{nh}$ where the *non-holomorphic* part T_{--}^{nh} has *no* holomorphic parts (i.e. functions of τ^- only), though it can have anti-holomorphic parts (i.e. functions of τ^+ only). The T_{--} (and likewise T_{++}) satisfies $\partial_+ T_{--} = -2\pi E \cdot \partial_- X$; in other words, T_{--} is conserved *on-shell* i.e. when the EOM $E^\mu = 0$ is satisfied. Hence for on-shell, $\partial_+ T_{--}^{nh} = 0$. This can be solved uniquely as T_{--}^{nh} has no holomorphic parts (unlike a similar equation for T_{--} which can not be so solved). In effect, this amounts to simply setting $T_{--}^{nh} = 0$ for on-shell T_{--} . All this can be checked carefully and there are no issues at this stage.

The next step involves a mode expansion of $G^\mu(\tau^-) = a \sum \alpha_m^\mu e^{-im\tau^-}$. This is also quite general. The Virasoro generators L_m follow in a straightforward way (though they are lengthy expressions). A curious feature of the L_m 's is that except for the terms coming from the quadratic part of the action S_0 , every oscillator appears contracted with e_+^μ . Therefore, apart from a normal ordering of the free parts, there are no factor ordering issues. This feature was also seen in the results of Aharony, Field, and Klinghofer, so there was really no need for any Weyl-ordering.

The next step was determining the quantum algebra of α_m^μ . Here we deviated from the standard path of deriving them from the Heisenberg commutation relations between Y^μ and their *canonical* momenta. The reasons for this were many but we cite one of the difficulties encountered; even at R^{-2} level, we saw an inconsistency as $\Delta[p^\mu, p^\nu] + \partial_{\sigma\sigma} \Delta[Y^\mu, Y^\nu] + \partial_\tau \Delta[Y^\mu, p^\nu]$ did not vanish (here $\Delta[A, B]$ is the deviation from the free field values). We ascribed this to the fact that the action was really of the *higher derivative* type for which no canonical formalism exists. Instead, one would have to use the famous Ostrogradsky formalism in its quantum version, further complicated by the presence of local invariances (see the summary in [30]).

Instead, we opted for bit of a trial and error approach to guess the oscillator algebra. The first observation was that the free oscillator algebra $[\alpha_m^\mu, \alpha_n^\nu] = m \eta^{\mu\nu} \delta_{m, -n}$ would not reproduce the Virasoro algebra. After some tedious work we did find an oscillator algebra that would correctly reproduce the Virasoro algebra with central charge $D + 12\beta$. Rather remarkably, one could find a redefinition of $G^\mu(\tau^-)$, equivalently a highly non-linear redefinition of the oscillator strengths, that reproduced the free oscillator algebra. The same oscillator redefinition also brought the $L_0, \bar{L}_0$

to their form that would obtain from S_0 . This meant that the spectrum was identical to that of the Nambu-Goto theory.

Two lacunae about this all order proof are (a) the string momentum was not analyzed to all orders, unless the field redefinitions adopted render all corrections vanish. This indeed happened with the field redefinitions used by [26], and, (b) the field redefinitions used are somehow not legitimate. As argued in [31], field redefinitions must respect what we had called *X-uniformity*.

Next, we come to the all order results for the Drummond type actions. Again one derives rather straightforwardly the EOM E^μ and the stress tensor T_{--} to all orders. It is noticed that every term in T_{--} for both of the Drummond type actions involves $D_{++}X$. This is shown to be non-holomorphic with the result that the entire T_{--} for Drummond type actions is non-holomorphic and does not contribute to the Virasoro generators. Because of this, no further attention was paid to the oscillator algebras. This part of the proof is straightforward and is unlikely to be wrong. The meaning of such actions which do not contribute to the Virasoro generators is something that needs to be explored further. Here too the string momentum was not analysed and the source of error could be there.

Finally, we come to the alleged claim of no corrections from any of the higher order terms arising out of the Covariant Calculus. While the proof is not as straightforward as in the case of Drummond type terms, a proof was given that the T_{--} from all of them are strictly non-holomorphic. Because of the rather detailed nature of these arguments, errors could well have been committed though that possibility is highly unlikely. It is important to stress that neither in ii) nor iii) field redefinitions and oscillator algebras played any role. But in both of them the string momentum issues were not examined.

Now we summarize the results of [24,26] and comment on their implications for the all order results. They found that even to order R^{-5} the ground state energy was the same as that given by the Arvis formula, but the energy of the first excited differed from that of the NG-theory, even though the deviations were still universal (see next section). This shows that the claims of [27] about ground state energies is correct to this order but not the claim about excitation energies. This is in spite of explicit corrections to string momentum found in [26]. This points to the oscillator algebra proposed in [27] to be problematic. The question remains whether the ground state energy continues to be that of Arvis even beyond this order. In [25] it is only anticipated that corrections at R^{-7} order and beyond are non-universal, but no calculations have been carried out. Also, in calculations to this order and beyond, the universal contributions coming from the Polyakov-Liouville action have to be carefully disentangled before the claims of [27] can be properly assessed. Later, in §11 of this chapter, we shall discuss the path-integral calculations of [32] where, at least in the saddle point approximation, it appears as if the claims of [27] for the ground state are correct. In the same work, they also considered the Polyakov extrinsic curvature action [33] as the leading correction. They found, again within the saddle point approximation, that the Arvis ground state energy was not altered. This is again in conformity with the analysis in [29].

Thus there are aspects of the all order claims that seem correct, while some others, like excited state energies, that are clearly wrong. Also, at some stage effective string theories must necessarily correct the tachyonic instability of the Arvis result. This is because QCD is a perfectly unitary theory. A reassessment of the all order claims is clearly in order.

22.10 Other Important Issues

22.10.1 The Excited States

As in the case of the quantization of strings [13], it is important to investigate the spectrum of the excited states, in addition to the ground state energies. Even before the quantisation of strings in [13], the framework for analysing spectrum was already there in the Dual Resonance model as explained in detail in Chap. 16. The essential aspects of a spectrum are the energies(masses) and their degeneracies. Group theoretical techniques developed in the Dual Resonance models were more or less reemployed in analysing the string spectrum too.

We first briefly review the theoretical status. Many of the formulae have already appeared earlier but have been collected here again for ease of presentation. In the context of flux tubes, the relevant analysis was performed by Arvis [8] though the energies of excited states was only implicit in his work. Making it explicit is just to replace the expression for the Arvis potential $V_{arvis}(R)$ by

$$E_{n,arvis}(R) = \sqrt{\{\sigma^2 R^2 + 2\pi\sigma[-\frac{(D-2)}{24} + n]\}} \quad (22.142)$$

For later use, we truncate this also to order $\frac{1}{R^3}$ and rearrange the terms to write it as

$$E_{n,arvis}^{trunc}(R) = V_{arvis}^{trunc} + \frac{\pi}{R}n + \frac{\pi^2}{R^3} \cdot \frac{1}{2\sigma}n(\frac{(D-2)}{12} - n) \quad (22.143)$$

Therefore, in the leading order (i.e. R^{-1}) one has uniform splitting of $\frac{\pi}{R}$ between excited states. The states at level n are all degenerate to all orders in R^{-1} .

But as already discussed this was valid only for $D = 26$ and the PS-effective string approach gave the same results at order R^{-1} but valid in all D . Their analysis was only restricted to R^{-1} order and they found agreement with the Arvis expression but now valid for all D .

The next thorough analysis of the spectrum of effective string theories was undertaken by Lüscher and Weisz in their 2004 work [3]. Their work was in the static gauge. They gave a very detailed analysis of the spectrum employing powerful group theoretical techniques and those of partition functions. While in $D = 3$ they found agreement for the Arvis spectrum truncated to R^{-3} , in D higher than 3 they obtained the general result for $n \leq 3$ (see their Eq.(6.1)):

$$\Delta E_{n,i} = \frac{\pi^2}{R^3} \left\{ n \left[\frac{(D-2)}{12} - n \right] c_2 + \nu_{n,i} (c_3 + 2c_2) \right\} \quad (22.144)$$

As already discussed they found one linear constraint between their c_2 and c_3 on the basis of their open-closed string duality, but this still left one free parameter. However, on the basis of a classical analysis of the Nambu-Goto action, we have already shown how both c_2 and c_3 get determined and on substitution of those values the above formula of LW coincides with that of truncated Arvis spectrum. The lack of degeneracy found by them disappears whenever $c_3 + 2c_2$ vanishes, as indeed happens for the values coming from the Nambu-Goto action.

The issue really hinged on whether there were any potential action terms at order R^{-3} over and above that of the Nambu-Goto term. This was settled by Drummond and us using the Polchinski-Strominger conformal gauge calculations. The upshot was that at even R^{-3} order the spectrum agreed with the truncated Arvis spectrum in all dimensions.

Finally, in a remarkable series of works Aharony and collaborators carried the analysis to order R^{-5} in both the static gauge and the conformal gauge. For the (1, 1) excited states they reported a correction to the Arvis spectrum, now truncated at R^{-5} level, of the form

$$\Delta E_{1,1;i} = -\frac{\pi^3 (D-26)}{48\sigma^2 R^5} C_{1,i} \quad (22.145)$$

where $C_{1,i}$ are coefficients that vanish in $D = 3$. They are also tabulated in Table 3 of Bastian Brandt's paper [34] which we will come to while reviewing the numerical results for the spectrum. A few comments are in order about this result. It is still a universal correction in that it only depends on D but through the combination $(D-26)$ in contrast to the Arvis spectrum which only involves $(D-2)$, the number of transverse degrees of freedom, to all orders. It also breaks the degeneracy of the Arvis spectrum.

Now we turn to some remarks about the numerical simulation results for the spectrum. One of the earliest numerical studies was by Juge, Kuti, and Morningstar in 2003 [35]. They had then reported deviations from the Nambu-Goto (Arvis) spectrum. The accuracies in those days could not have probed the R^{-5} levels where deviations are expected on theoretical grounds. Most likely these deviations were systematic errors.

Bastian Brandt has very carefully investigated the spectrum but only in $D = 3$ where the R^{-5} corrections are absent [34]. He has used large Wilson loops as well as Polyakov loop correlators in his study as well as powerful techniques to control various sources of contamination and systematic effects. He concludes that while there is qualitative agreement between the numerical results and the Arvis spectrum, there are some quantitative deviations, particularly at large values of R . But large values of R are also numerically very demanding. If these deviations survive further scrutiny and reduction of systematic errors, there will certainly be reasons to revisit the effective string theories themselves.

22.10.2 AdS-CFT Approaches

The so called AdS-CFT correspondence, also called the Gauge-Gravity duality [36, 37], is one of the most remarkable of connections ever made. There is an ever increasing literature on this and it is way beyond the scope of this book to even attempt a heuristic discussion. We refer the reader to the course on this subject by Benini [38] for a good introduction, which also has references to important sources. We shall be content with making some broad comments in so far as it directly concerns one of the essentials of our book, namely, the static potential between quarks and anti-quarks.

Broadly speaking, this correspondence in a rather precise sense is between gravity theories in $d + 1$ dimensions and Conformal Field Theories (CFT) on the d -dimensional boundary. More specifically, the gravity theories considered are on Anti-de Sitter (AdS) backgrounds. Benini gives a simple motivation that is particularly relevant to our book. As we have already discussed, string theories can be made sensible even for $D \leq 26$ (the so called sub-critical theories) at the cost of introducing an additional degree of freedom i.e. the liouville field. This can be thought of as providing an additional ‘dimension’ z and consider a metric on the enlarged space of the form

$$ds^2 = w(z)^2(dx_D^2 + dz^2) \quad (22.146)$$

Now if one demands scale invariance wrt to the original coordinates i.e. $x \rightarrow \lambda x$, it can be realized provided the new coordinate is likewise scaled i.e. $z \rightarrow \lambda z$ and the scale factor (often called the *warp factor*) is chosen to be $w(z) = \frac{R}{z}$ so that

$$ds^2 = R^2 \frac{dx_D^2 + dz^2}{z^2} \quad (22.147)$$

which is an AdS metric.

What is of interest in the context of our book is the application of the correspondence to Wilson loops $W(\mathcal{C})$ defined around a loop $\mathcal{C}$. More precisely, it is to find the gravity analog for the gauge theory observable. The prescription is to take Nambu-Goto action for AdS and find the minimal area surface that is bounded by $\mathcal{C}$. When only the classical AdS geometry is used, it should lead to the linearly rising confining potential. The idea was then put forward to take into account various quantum fluctuations to probe corrections to σR term; in particular the Lüscher term. Naik [39] showed that for $D = 3$ the Lüscher term with the correct sign could be reproduced by only taking into account fluctuations in the radial AdS coordinate. Greensite and Olesen [40] considered world-sheet fluctuations in a general supergravity background and found the Lüscher term with the wrong sign casting doubts on the utility of supergravity calculations for the observed Lüscher term in lattice simulations of QCD (see the detailed discussion of this in Chap. 21).

We refer the reader to these works for details. Additionally, Förste et al. [41] also investigated the implications of the AdS-CFT correspondence for Wilson loops.

While all these works were able to reproduce the Lüscher term, they have nowhere reached the levels the effective string theory descriptions have reached. In fact, even the absence of corrections at R^{-2} level have not been reproduced by the AdS-CFT approaches. Like in all approaches, higher order calculations in AdS-CFT become increasingly unwieldy both technically as well as conceptually. But these must be done to complement what has been learnt from effective string theories so far. It is also not clear how the AdS-CFT approach will account for the non-universal corrections at R^{-7} order.

22.10.3 Thickness of Flux Tubes

One of the striking aspects of the simulations by Bali et al. [42, 43] and Haymaker et al [44], discussed in Chap. 21 is that the flux tubes appear to have noticeable *thickness* (see also [45, 46]). This is clear from both the action and energy profiles of Bali et al. A precise interpretation of these broadened profiles is not so straightforward. From an effective string point of view it could arise as both due to the fluctuations of a very narrow string and/or an intrinsic thickness to the effective string. It is also not clear how to disentangle these two aspects.

The topic of thickness of flux tubes has been of interest for a very long time. Already in 1981 Lüscher, Berg, and Weisz had investigated this in [47] and had argued for a logarithmic increase in thickness with separation. Such an increase can not be ascribed to any intrinsic thickness.

The broadening of QCD flux tubes in $D = 3$ has also been addressed from an AdS-CFT correspondence by Greensite and Olesen [48]. Polchinski and Susskind [49] made the intriguing suggestion that four-dimensional projection of certain *thin* AdS_5 strings behave like thick strings. Vyas [50] has further elaborated this point of view. In particular, he has stressed the importance of *massive* modes in this context. How to include such massive modes in effective string descriptions is far from clear. From one point of view, integrating out such massive modes would again give effective string theories of the kind we have already considered. A resolution may be in examining observables over and above those of the spectrum. Needless to say, this important issue requires further studies. Another idea that may turn out to be fruitful in this regard is that of *Fat* strings in AdS. For this interesting perspective see the article by Das [51]. The bulk of our book has dealt with the theme of going from hadronic strings to fundamental strings; in [51] Sumit Das explores how the journey can happen the other way too i.e. from fundamental strings to QCD strings.

It is also very important to carry out the spectrum studies in $D = 4$ where the R^{-5} correction first shows up. While the $SU(3)$ case was already very challenging even at the level of R^{-3} (see [4, 5]), going to R^{-5} level may just be too difficult. But a study of $SU(2)$ or even Z_2 may already throw important light on these issues.

22.11 Path Integral Quantization of Subcritical Strings

We now discuss approaches based on the path integral quantization of the Polyakov action of Eq. (22.48) and possible additions to it. We shall mainly be focussing on the works of Durhuus et al. [52] on dual models in the saddle point approximation, of Durhuus et al. [53] on the static potential in the Polyakov theory, and finally, of Ambjorn et al. [32] on effective QCD strings in the Polyakov approach. Though chronologically the first two papers appeared well before the developments elaborated in the earlier sections of this chapter, we are discussing them the last in order to contrast them with the later developments. Though the Lüscher-Weisz approaches were also based on path-integral methods, they were done in the static gauge for essentially the Nambu-Goto like theories.

Polyakov, in his pioneering work [15], had stated that at a classical level, his action is completely equivalent to the Nambu-Goto action. This could be seen on noting that the saddle point solution, which also happens to be exact, simply equates the intrinsic metric to the induced metric(modulo an irrelevant local scale factor). But at the quantum level, there are dramatic differences. Polyakov had himself demonstrated one of these in [15]; that was that his theory could be formulated consistently in dimensions other than $D = 26$. In particular, the appearance of the Liouville action for the so called sub-critical dimensions i.e. $D < 26$.

The works to be discussed now [32,52,53] all follow the path integral approach to quantization, as was also done by Polyakov in [15]. A substantial improvement over Polyakov's original work was the incorporation of boundaries. At the level of actions, open strings need boundary terms while close strings do not. But at the level of the world sheets, both of them have to be described by manifolds with boundary. For open strings this manifold can be taken as the one with disk topology with rectangular boundaries, while for closed strings the manifolds are of cylinder topology. Manifolds with boundaries bring in a number of subtleties, and the behaviour of the Liouville field at the boundaries is essential for the saddle point solutions to exist. All these works, at one point or the other, invoke large $|D|$, saddle point techniques, and the so called mean field solutions(which are also based on large $|D|$). All of them express the hope, without sufficient justifications, that even the large $|D|$ solutions will somehow be relevant even for real-life cases of $D = 4, 3$. This certainly can not be the case when comparisons are made to high accuracy numerical simulations, where even differences between $D = 3$ and $D = 4$ are clearly noticeable.

In [52], the authors study the path-integral formulation of Polyakov's theory in the saddle point approximation. They further restrict their analysis to the conformal gauge for the intrinsic metric. They only carry out the X^μ -integrations. They find that a saddle point solution can exist in the $D \rightarrow -\infty$ limit, and that too provided the intrinsic metric is singular on the boundary. Proceeding in accordance with Polyakov's proposal they are able to construct off-shell Green functions for the scattering of spin-0 mesons. They neglect the Faddeev-Popov determinant (which gives rise to the contribution proportional to -26 in the Liouville action) as they are only dealing with the $D \rightarrow -\infty$ case. They also find that when the external masses obey a certain condition, the functional integration over the Liouville field completely

decouples. Subsequently they find that the on-shell S-matrix elements reproduce the Veneziano amplitudes. They offer no explanations for why a $D \rightarrow -\infty$ calculation should do so, nor any indications of how corrections to the saddle point solution are to be treated systematically.

In [53], the authors study the static potential, of central interest to this book, in the Polyakov theory. They too resort to saddle point techniques in the $D \rightarrow -\infty$ limit. In particular, they study large Wilson loops. The boundary of the manifold is mapped to the Wilson loop. In addition to the Polyakov action, they consider (for renormalization reasons) the action

$$\mu_0^2 \int_D d^2z \sqrt{g} \quad (22.148)$$

where D is the world sheet domain. It should be noted that they use the notation h_{ab} for the induced metric and g_{ab} for the intrinsic metric! This term, which is not Weyl-invariant, leads to a term of the type $\mu_0^2 e^\phi$ in the Liouville action. In our considerations of effective string theories, such a term has not been considered. In addition, they also include a boundary action.

The authors invoke what they call a “double saddle-point” approximation; the first of these realizes the $R \rightarrow \infty$ limit, where R is the quark-antiquark separation, and the second realizes the $D \rightarrow -\infty$ limit. They too perform the calculations in the conformal gauge for the intrinsic metric and the saddle point is for the combined X^μ and Liouville fields. The functional integrals are first carried out with the Liouville field fixed both in the interior and on the boundary (where it takes the value $\psi(z)$). This leads to the Liouville action (proportional to $D - 26$). The saddle point approximations are made while carrying out the functional integrations over the fields $\phi(z)$, $\psi(z)$. Though they investigated the static potential for large R , they only determined the linearly rising and R^{-1} pieces. When $\mu \neq 0$ they reached the surprising conclusion that the R^{-1} term actually vanishes (actually, they could not evaluate the saddle point exactly, and had to make some approximations)! It is important to understand the meaning of this result, if correct, in the light of our earlier discussions of the Lüscher term. For $\mu = 0$ they obtain the R^{-1} correction to the linearly-rising part of the potential to be $-\frac{\pi D}{24R}$ whereas the universal Lüscher term is $-\frac{(D-2)\pi}{24R}$. But as the authors have clearly clarified, their large D approach can only capture the part proportional to D . This clearly points to the need for understanding, in a systematic manner, corrections to their large D saddle point approximations. Large D techniques and $\frac{1}{D}$ corrections to them were pioneered by Alvarez [54,55].

Now we come to the work of Ambjorn et al. [32] which addresses effective QCD strings beyond Nambu-Goto action. They too choose to use the path-integral quantization of Polyakov theory for world sheets with boundaries mapped to the Wilson loops. This is what one would have done had the strings in question been fundamental. This is in contrast to the way effective strings had been dealt with in the earlier parts of this chapter i.e. treat them as being valid only for small fluctuations around the classical string configuration, and treat the fluctuations quantum-mechanically in perturbation theory.

They too work in the conformal gauge for the intrinsic metric, so that the Liouville field $\phi(z)$ is the only remaining degree of freedom of the intrinsic metric. Many technical features are akin to those of [53]. They treat the path-integral as if the full fluctuations of X^μ -field are relevant (not just small fluctuations around the classical string configuration). On denoting the boundary value of $\phi(z)$ (the Liouville field) by $\psi(z)$, they solve the EOM for X^μ with ψ fixed and denote the solutions by $X_\mu^\psi(z)$. They integrate over the fluctuations of X^μ around X_μ^ψ (being just a quadratic functional integration), recovering the Liouville action of Polyakov.

What is left are functional integrations over the Liouville field $\phi(z)$ in the interior as well as $\psi(z)$ on the boundary. They simplify matters using the elegant Upper Half Plane (UHP) parametrization. Nevertheless, they are able to perform the functional integration only in the $D \rightarrow \infty$ and that too using the saddle point approximation. In this limit, the so called mean-field approximation becomes exact. These authors too make the claim, without adequate justification, that for the Polyakov formulation of the Nambu-Goto string, the mean field result is exact even for finite D . Till subleading (in D) corrections are actually calculated (as in [55]) and shown to vanish, one should keep the fingers crossed. Though such a possibility sounds extraordinary, it is still conceivable; the Duistermaat-Heckman theorem [56] about conditions for exactness of WKB approximations is a case in point. We refer the reader to the many interesting details in [32] and simply state their salient results.

They find the static potential to match exactly i.e. to all orders in R^{-1} , the Arvis result. As discussed extensively in Sect. 22.9 of this chapter, this was also the claim made by [27]. If the saddle point (mean field) results do not indeed get corrected, these are strong vindications of [27] (see Sect. 22.9 for detailed comparisons). Rather remarkably, this path-integral approach appears to bypass the thorny issues of string momentum, oscillator algebras etc., inherent to the operator methods. As is well known, detailed comparisons between path-integral and operator quantizations are notoriously hard. As discussed earlier, Aharony and collaborators found deviations from the Nambu-Goto theory for the energies of the first excited states at order R^{-5} . Unfortunately, only the ground state energies were addressed in [32]. Their calculations should be extended to the full spectrum, at least to the first excited states.

The authors attempt the simplest generalization of the Polyakov action by adding an extrinsic curvature action, first proposed by Polyakov in [33]. The form of the extrinsic curvature action used by them is

$$S_{extr} = \frac{1}{2\alpha} \int d^2z \sqrt{g} \Delta X \cdot \Delta X \quad (22.149)$$

(It should be recalled that in their notation g is the intrinsic metric). Here α is a dimensionless constant and Δ is the 2d Laplace-Beltrami operator. However, the leading order corrections to the effective action proposed by Drummond as well as both our covariant calculi, discussed earlier, are not quite of this form. In fact we had proposed two independent leading order corrections. One has to check whether there is any equivalence of one of those, or a linear combination of those, modulo

EOM's and constraints of previous order, to Polyakov's extrinsic curvature action in the specific form used by the authors. With the inclusion of the extrinsic curvature action, the saddle point analysis also gets considerably harder. The authors again resort to mean field analysis. They find that for large R , the extrinsic curvature terms do not alter the Arvis result for the ground state. As already commented in §9, this too is in conformity with the claims made in [28,29]. Once again, it is of utmost importance to calculate the corrections to the mean field results. However, an earlier calculation by Braaten and coworkers [57], also using large D analysis, found an R^{-4} correction. Around the same time as them Olesen and Yang [58] had also analysed the extrinsic curvature action, again in the large D limit and saddle point approximation. They specifically studied the static potential. As already emphasized, the saddle point analysis even in the large D limit is quite difficult. Olesen and Yang made many further approximations which are not very systematic. They reported results for the R^{-1} correction in two distinct circumstances: (a) non-perturbative, in the sense of not doing a perturbative analysis for small fluctuations - they claim a deviation of the coefficient from the universal Lüscher value. The high accuracy numerical data discussed in Chap. 21 do not seem to support such a deviation, but a reanalysis of the data with these specific non-perturbative corrections in mind may be worthwhile, and (b) perturbative analysis - here they find the D -dependent piece of the standard Lüscher term. Once again, a thorough reanalysis of the extrinsic curvature actions is in order.

22.12 Concluding Remarks

In this book we have narrated, with necessary technical and conceptual elaborations, how strings made an appearance as the logical culmination of a number of deep, fascinating and powerful attempts at understanding strong interactions. These were the ideas of S-matrix, Dispersion Relations, Complex Angular Momenta and Regge Poles, Duality and the Dual Resonance Models. We have also explained the many conceptual problems that arose with the string theoretic description of strong interactions, and how they were replaced by a relativistic quantum field theory of quarks and gluons called Quantum Chromodynamics. The non-observance of quarks led to a crisis for this route to strong interactions too, but the ingenious proposal of the dual superconductor mechanism pointed to a way out, at least in principle.

A thin flux tube was at the heart of the matter. We have then described in detail in the last two chapters how this flux tube, established through impressive numerical simulations based on Lattice Gauge Theories, not just superficially resembles a string but to a very high degree of accuracy (more specifically, the first six terms in the large- R expansion of the static quark-antiquark potential) resembles a Bosonic String Theory even mathematically.

Never before in the history of physics has there been a situation where two radically different theories for the same physical phenomena have eventually reached a conceptual fusion. This is what motivated the author to call this fantastic occurrence as *Strings to Strings*!

A central theme to the analytical understanding of the flux tube has been that of effective string theories. In spirit as well as in their mathematical implementation, these are like other effective descriptions which have been thoroughly treated in Chap. 18. For strong interactions, such effective descriptions were the celebrated Chiral Effective Field Theories. While in their case, the symmetry content could be more easily understood from the point of view of the microscopic theory, QCD, in the case of effective string theories, even this is rather obscure. That is, what is a microscopic understanding of the world-sheet general coordinate invariance from a QCD point of view?

Hopefully, the detailed picture of flux tubes given by effective string theories, albeit in a truncated version of QCD without dynamical quarks, will shed light on these fundamental issues. Analytical approaches to flux tubes as pioneered by Adler [59] or Baker et al. [60], or the more recent ones of Akhmedov et al. [61] should be revived in the light of the progress made on the effective string theory front.

Despite half a century since the inception of QCD we still do not have first principles understanding of several issues like the pion mass, pion-nucleon coupling constant, confinement etc. It is not enough for any understanding of confinement to just produce the linearly rising confining part; it should also reproduce the sub-leading terms. Both numerical simulations as well as effective string theory calculations should be pushed to probe even higher order corrections to shed more light on the non-universal terms in the static potential. It is obvious that such non-universality has to be present in the static potential as at shorter distances it is governed more by asymptotic freedom which is sensitive to non-universal features like the color group, quark representations etc. More specifically, a much better understanding of the region in between the string-dominated region on the one hand, and asymptotic freedom dominated region on the other, is desirable.

Ironically, the very S-matrix program that fuelled the important developments in strong interactions itself becomes fuzzy as the asymptotic states of QCD are not the quark and gluon states. So a revision of the tenets of S-matrix theory in the light of QCD is called for. On the effective string theory side, the ideas and techniques developed by Simeone Hellerman and his collaborators [62] should be more effectively (pun intended) integrated into the overall framework of effective string theories. The precise connections between calculations in different gauges needs to be put on a stronger footing. We also did not elaborate the Polyakov approach based on the intrinsic metric much.

One of the eventual hopes is to be able to explain as much of hadronic physics as possible in terms of a weakly interacting string theory.

References

1. M. Lüscher, P. Weisz, JHEP **0207**, 049 (2002)
2. M. Lüscher, P. Weisz, JHEP **09**, 010 (2001)
3. M. Lüscher, P. Weisz, JHEP **0407**, 014 (2004)
4. N.D. Hari Dass, P. Majumdar, PoS(Lattice2005), p. 312
5. N.D. Hari Dass, P. Majumdar, JHEP **0610**, 020 (2006)

6. N.D. Hari Dass, P. Majumdar, PoS LATTICE 2007 p. 316 (2007)
7. N.D. Hari Dass, P. Majumdar, Phys. Lett. B **658**, 273–278 (2008)
8. J.F. Arvis, Phys. Letts. **B127**, 106 (1983)
9. Y. Nambu, Phys. Lett. B **80**, 372 (1979)
10. K. Dietz, J. Filk, Phys. Rev. D **27**, 2944 (1983)
11. J. Polchinski, *String Theory*, vol. I, Cambridge University Press
12. J. Ambjorn, P. Olesen, C. Petersen, Phys. Lett. B **142**, 410 (1984)
13. P. Goddard, J. Goldstone, C. Rebbi, C.B. Thorn, Nuc. Phys. **B56**, 109 (1973)
14. J. Polchinski, A. Strominger, Phys. Rev. Lett. **67**, 1681 (1991)
15. A.M. Polyakov, Phys. Lett. B **103**, 207 (1981)
16. K. Fujikawa, Phys. Rev. Lett. **42**, 1195 (1979)
17. J.M. Drummond, hep-th/0411017
18. N.D. Hari Dass, P. Matlock, hep-th/0606265
19. J. Drummond, hep-th/0608109v1
20. N.D. Hari Dass, P. Matlock, *Covariant calculus for effective string theories*, hep-th/0709.1765
21. N.D. Hari Dass, P. Matlock, Covariant calculus for effective string theories. Ind. J. Phys. **88**, 965–977 (2014)
22. N.D. Hari Dass, *On string momentum in effective string theories*, hep-th/1005.4829v1
23. S. Weinberg, *Gravitation and Cosmology* (Wiley, 1972)
24. O. Aharony, E. Karzbrun, JHEP **0906**, 012 (2009)
25. O. Aharony, Z. Komargodski, Effective theory of long strings. JHEP **05** (2013)
26. O. Aharony, M. Field, N. Klinghofer, JHEP **04**, 048 (2012)
27. N.D. Hari Dass, P. Matlock, Y. Bharadwaj, *Spectrum to all orders of Polchinski-Strominger Effective String Theory of Polyakov-Liouville Type*, hep-th [arXiv:0910.5615](https://arxiv.org/abs/0910.5615)
28. N.D. Hari Dass, Y. Bharadwaj, *Spectrum to all orders of Polchinski-Strominger Effective String Theories of Drummond Type*, hep-th [arXiv:0910.5620](https://arxiv.org/abs/0910.5620)
29. N.D. Hari Dass, *All Conformal Effective String Theories are Isospectral to Nambu-Goto Theory*, hep-th [arXiv:0911.3236](https://arxiv.org/abs/0911.3236)
30. N.D. Hari Dass, AIP Conference (IX QCHS, Madrid 2010) Proceedings No. 1343 (2010), p. 230–232
31. N.D. Hari Dass, P. Matlock, *Field Definitions, Spectrum and Universality in Effective String Theories*, hep-th [arXiv:061229v1](https://arxiv.org/abs/061229v1)
32. J. Ambjorn, Y. Makeenko, A. Sedrakyan, Phys. Rev. D **89**, 10 (2014)
33. A.M. Polyakov, Nucl. Phys. **B268**, 406 (1986)
34. B.B. Brandt, Indian J. Phys. **95**(8), 1613 (2021)
35. K.J. Juge, J. Kuti, C.J. Morningstar, Phys. Rev. Lett. **90**, 161601 (2003). (J. Kuti, Proceedings of Lattice 2005, PoS(Lattice 2005), 1)
36. V. Vyas, Phys. Rev. D **87**, 4 (2013)
37. V. Vyas, *Flux Tubes in Confining Gauge Theories with Gravitational Dual*, hep-th/1904.06777
38. F. Benini, *Introduction to AdS/CFT*
39. S. Naik, Phys. Lett. B **464**, 73 (1999)
40. J. Greensite, P. Olesen, JHEP **04**, 001 (1999)
41. S. Förste, D. Ghoshal, S. Thiesen, JHEP **08**, 013 (1999)
42. G.S. Bali, C. Schlichter, K. Schilling, Phys. Rev. D **51**, 5165 (1995)
43. G.S. Bali, Quark forces and heavy quark bound states. Phys. Rep. **343** (2001)
44. R.W. Haymaker, V. Singh, Y. Peng, J. Wosiek, Phys. Rev. D **53**, 389 (1996)
45. M. Casselle, M. Panero, D. Vadachin, JHEP **02**, 180 (2016)
46. M. Casselle, P. Grinza, JHEP **11**, 174 (2012)
47. M.Lüscher, B. Berg, P. Weisz, Nucl. Phys. **180**, 1 (1981)
48. J. Greensite, P. Olesen, JHEP **11**, 030 (2000)
49. J. Polchinski, L. Susskind, *String Theory and Size of Hadrons*, hep-th/0112204
50. V. Vyas, *Intrinsic Thickness of QCD Flux Tubes*, hep-th/1004.2679
51. S. Das, *QCD, Strings and Emergent Space*. Eur. Phys. J. ST. **231** (2022)
52. B. Durhuus, H.B. Nielsen, P. Olesen, J.L. Petersen, Nucl. Phys. **B196**, 498 (1982)

-
53. B. Durhuus, P. Olesen, J.L. Petersen, Nucl. Phys. **B232**, 291 (1984)
 54. O. Alvarez, Phys. Rev. **24**, 440 (1981)
 55. O. Alvarez, Nucl. Phys. **B216**, 125 (1983)
 56. J.J. Duistermaat, G.J. Heckman, Invent. Math. **69**(2), 259(1982)
 57. E. Braaten, R.D. Pisarski, S.H. Tze, Phys. Rev. Lett. **58**, 93 (1987)
 58. P. Olesen, S.-K. Yang, Nucl. Phys. **B283**, 73 (1987)
 59. S.L. Adler, Nucl. Phys. **B217**, 381 (1983)
 60. M. Baker, J.S. Ball, F. Zachariasen, Phys. Rev. D **34**, 3894 (1986). (M. Baker, J.S. Ball, F. Zachariasen, Phys. Rep. **209**, 73 (1991), M. Baker, J.S. Ball, F. Zachariasen, Phys. Rev. D **41**, 2612 (1990))
 61. E.T. Akhmedov, M.N. Chernodub, M.I. Polykarpov, M.A. Zubkov, Phys. Rev. D **53**, 2087 (1996)
 62. S. Hellerman, S. Maeda, *On vertex operators in effective string theories*, [arXiv:1701.06406](https://arxiv.org/abs/1701.06406) (hep-th). (S. Hellerman, I. Swanson, *Boundary Operators in effective string theories*, JHEP 04 085(2017))

Index

Symbols

α , β rays, 19
 β -decay theory, 22
 β -electrons, 19
 continuous spectrum, 20
 line spectrum, 20
 spectrum, 20
 γ -rays, 19
 π N interactions, 37
 equivalence of pseudoscalar and
 pseudovector, 37
 $\pi^\pm$ -decays, 292
 π_0 decay, 37
 ρ -trajectory, 197
 σ -model, 299
 massless pions, 299
 nucleon mass, 299
 spontaneously broken, 299
(3, 3) resonance, 164
5 point function, 214
 detailed construction, 216
 duality, 215
 Y.Nambu, 218
6 point functions, 220
 duality, 220
 explicit construction, 220

A

Abelian Higgs model, 346
Absorptive part, 135, 150
 observable, 140
Action without Hamiltonian, 67
Adler consistency conditions, 306
Against self-energy, 67

Against vacuum polarization, 70
All order calculations, 74
 relevance?, 74
Altarelli-Parisi equations, 342
Analyticity
 causality, 127
 in masses, 124, 153
Analyticity domains, 152
 in invariants, 153
 in masses, 154
 maximal, 154
 off shell, 154
Anderson–Higgs mechanism, 25
Anomalies, 311
 π^0 -decay, 311
 SU(2) $\times$ SU(2) effective actions, 311
 Wess–Zumino actions, 311
Anomalous dimensions, 333
Anomalous magnetic moment, 65
Anomaly, 243
Antiparticles, 7
A RQFT Primer, 77
Arvis quantization, 272
 boundary conditions, 272
 changed mode expansions, 272
 constraints, 272
 full gauge fixing, 272
 non-covariant approach, 272
 normal ordering, 274
 rotational invariance, 274
 strings with fixed ends, 272
 tachyonic instability, 274
 transverse d.o.f, 273
A Sad Theorem, 198

- Asymptotic conditions
 - Dyson S-matrix, 93
 - strong, 86
 - weak, 81, 97
- Asymptotic freedom, 330
 - ϕ^3 in D-6, 342
 - Coleman-Gross theorem, 342
 - non-abelian gauge theories, 342
 - QCD, 339
- Atomic mass, 6
- Atomic number, 6
- Axial vector currents, 293
 - exact conservation, 293
 - σ -model, 297
 - σ -particle, 296
 - PCAC, 293
- B**
- Baker-Campbell-Hausdorff Identity (BCH), 234
- Bakerian lectures, 7
- BEG analyticity, 167
- Beta functions, 330, 333
 - gauge independence, 334
 - negativity in QCD, 338
 - QCD leading, 338
 - scheme dependence, 334
- Bjorken scaling, 325
- Bohr–Sommerfeld model, 7
- Bootstrap conditions, 212
 - unwanted daughters, 212
- Bose–Einstein statistics, 10
- Bosons, 10
- Bottled neutrons, 280
- Bottom quark, 27
- Bound states, 75
- Box normalization, 83
- C**
- Cabibbo angle, 280
- Callan-Symanzik equations, 330
- Carlson’s Theorem, 178
 - elastic unitarity, 185
- Causality
 - and antiparticles, 56
 - Coulomb Gauge QED, 53
 - from dispersion relations, 129
 - philosophers, 56
- Causality and analyticity, 46
 - classical electrodynamics, 51
 - NRQM, 52
 - QED, 59
 - RQFT, 52
- Causality and Chance, 56
- Causality conditions, 52
- Causal transforms, 129
- Central extension, 243
- Charged pion decay, 164
 - dispersion relations, 164
 - Goldberger-Treiman relation, 164
- Charged pions, 35
 - exptal discovery, 35
 - spin, 36
- Charm quarks, 26
 - discovery, 26
- Chew-Frautschi plot, 186
- Chew-Low theory, 37
- Chiral cancellations, 302
- Chirality non-conservation, 170
- Chiral perturbation theory, 308
- Chiral symmetry, 297, 323, 347
 - breaking by quark masses, 347
 - lattice QCD, 348
 - non-linear realization, 301
 - QCD, 347
 - spontaneous breaking, 297, 347
- Chiral transformations, 296
 - covariant derivatives, 302
 - group structure, 297
- Classical strings, 258
 - action, 258
- Coherent states, 235
 - as eigenstates, 235
 - evolution, 235
 - important properties, 235
 - overlap, 235
 - parameter shift, 235
- Colour, 319
 - broken in HN, 323
 - exact, 321
 - Gell-Mann Zweig, 321
 - Han-Nambu, 319
 - π^0 -decay, 321
 - statistics, 321
- Complex angular momentum
 - conditions for, 179
 - dispersion relations, 179
 - going beyond Lehmann Ellipses, 178
 - and Mandelstam rep., 179
 - potential scattering, 179
- Complex frequencies, 47
- Confinement, 390
 - coexistence with AF, 390
 - and LGT, 390
- Continuum limit, 360
 - naive, 360
 - statistical, 360
- Cooper Pairs, 346

- Coulomb scattering, 65, 67
- Covariant calculus, 424
 - conformal gauge I, 431
 - conformal gauge-II, 432
 - drummond actions, 435
 - equivalence, 434
 - gauge fixing, 431
 - Nambu Goto way, 425
 - Polyakov way, 427
 - static gauge, 431
- Covariant derivatives, 427
 - Leibnitz rule, 429
 - reparametrisation, 427
 - Weyl, 428
 - Weyl-reparametrisation, 427
- Covariant Schrödinger Eqn., 66, 71
- Criticism of Kronig, 9
- Critique of Salam's dispersion relations, 150
- Crossing symmetry, 49
 - Jackson's error, 49
 - in optics, 49
- Current algebra, 169, 190
- Current-current interaction, 22
- CVC hypothesis, 280, 292

- D**
- Dalton's law of proportions, 4
- Dalton's laws, 4
- Daughter trajectories, 198
- D-branes, 262
- Deep inelastic scattering, 325, 341
 - scaling, 341
- Dielectric, 51
 - dispersion, 51
- Dimensional regularization, 334
- Dimensional transmutation, 339
- Dirac and the path integrals, 69
- Dirac brackets, 265
 - solving the constraints, 265
- Dirac Eqn., 7, 12, 60
 - Dirac Sea, 60
 - discrepancies, 61
 - negative energy solns., 60
 - spin, 7, 12
- Dirac Sea, 60
 - holes as protons, 60
- Discrete bound states, 137
- Dispersion, 46
 - Druid, 45
 - Heisenberg-Kramers, 46
 - Lorentz, 45
 - vs dispersion relations, 46
- Dispersion relations, 45, 90, 108
 - analyticity, 94
 - and S-matrix, 127
 - causality, 46
 - crossing symmetry, 104
 - details of, 47
 - fixed-t, 108, 144, 151
 - forward scattering, 108, 132
 - Kramers-Kronig, 104, 127
 - limitations, 189
 - negative frequencies, 49
 - number of subtractions, 180, 182
 - off shell, 157
 - optics, 47
 - in QFT, 130
 - relating to observables, 134
 - RQFT, 127
 - Salam's approach, 145
 - subtractions, 134
 - unsubtracted, 140
- Dispersive part, 135, 150
- Displacement operator, 84
- Divergences, 63
 - infrared, 62
 - quadratic for spin-0, 63
 - self-energy, 63
 - ultraviolet, 62
- Divergence types, 73
- Domains of analyticity, 143
 - Symanzik analysis, 143
- Double spectral functions, 159
 - π N-scattering, 159
 - support, 159
- Double spectral rep., 159
 - fixed-t dispersion relations, 159
 - perturbation theory, 159
 - QED, 159
- Double spectral representation, 159
- Duality, 127, 194, 351
 - Dolen-Horn-Schmid, 194
 - generalized Ising models, 351
 - graphical methods, 214
 - Kramers-Wannier, 352
- Duality constraints, 223
 - Chan-Tsun soln., 223
 - Koba-Nielsen soln., 223
- Dual resonance
 - ghosts, 244
- Dual resonance models, 233
 - absence of ghosts, 248
 - critical dimension, 248
 - DDF construction, 248
 - degeneracies of states, 241
 - elimination of ghosts, 246
 - factorizability, 233

- first two levels of excitation, 249
- operator formalism, 219, 233
- physical states, 244
- positive norm states, 244
- similarities to QED, 245
- spurious and null states, 247
- spurious states, 250
- transverse states, 246
- Virasoro Subsidiary conditions, 246
- zero-norm states, 250
- Dual superconductor mechanism, 345
- Dyson's Equivalence of QED theories, 62
- Dyson's equivalence proof, 71
- Dyson's QED S-matrix, 62
- E**
- Eden-Kinoshita bound, 168
- Effective field theories
 - effective Lagrangeans, 289
 - elasticity, 279
 - electroweak theory, 280
 - Fermi theory, 280
 - Ginzburg–Landau theory, 283
 - microscopic basis, 280
 - σ -model, 290
 - source theory, 289
 - strong interactions, 288
 - superconductivity, 283
 - symmetries, 280
 - $V - A$ theory, 280
 - weak interactions, 280
- Effective Lagrangeans
 - anomalous sector, 310
- Effective string theories
 - AdS-CFT, 443
 - beyond Nambu-Goto, 446
 - excited states, 441
 - extrinsic curvature, 440
 - Lüscher-Weisz, 403
 - path integral approach, 445
 - and QCD, 448
 - R^{-5} potential, 437
 - R^{-5} spectrum, 437
 - static gauge, 403
 - thickness, 444
- Effective theories, 15
- Eightfold way, 317
 - Ω^- , 317
 - SU(3), 317
- Electric-Magnetic duality, 344
- Electron spin, 9
- Electroweak unification, 26
- Elementary vs composite, 45
- Empirical mass formula, 31
- Equivalence of field theories, 66
- Euclideanization, 355
 - case of gravity, 357
 - positivity of actions, 357
 - reflection positivity, 356
- Euclidean postulate, 355
 - nature's language, 355
- Euler Beta-function, 204
 - Veneziano formula, 204
- Euler-Mascheroni const. γ_E , 344
- F**
- Factorizability, 233, 255
 - oscillator structure, 233
- Fermi coupling constant, 22
- Fermi–Dirac statistics, 10
- Fermion doubling, 353
- Fermions, 10
- Feynman matrix element, 130, 146
- Feynman propagator, 89
- Feynman's particle perspective, 70
- Feynman-Wheeler theory, 69, 70
- Fields
 - are they necessary?, 77
 - asymptotic, 78
 - interpolating, 82
 - normalization, 82
 - vs particles, 77
- Fine structure, 12
- Finite Energy Sum Rules (FESR), 191, 193
 - criticisms, 198
 - recent applications, 201
- Finite pion mass, 303
- Flux tube profiles, 392
 - numerical studies, 392
 - thickness, 394
- Flux tubes
 - action density, 382
 - analytical studies, 378
 - and dual superconductors, 377
 - Creutz, 383
 - energy density, 382
 - in LGT, 377
 - LGT observables, 379
 - Lüscher term, 391
 - Monte Carlo simulations, 378
 - profile measurements, 381
 - profiles, 378
 - static potentials, 378
 - Wilson vs. Polyakov loops, 381
- Foliation, 71
- FORM, 340
- Forward scattering, 108, 132
 - details of massive case, 139

- dispersion relations, 108
- massive case, 134, 140
- massive vs massless cases, 108
- massless case, 132
- Symanzik analysis, 143
- Fritzsch-Gell-Mann colour, 324
- octet of gluons, 324
- Froissart bound
 - conformal field theories, 169
- Froissart bounds, 166
- Fubini-Furlan method, 171
- G**
- Gamow–Teller transitions, 22
- Gauge-gravity duality, 168
- Gauge invariance, 285
 - not a symmetry, 285
- Gell-Mann-Nishijima formula, 316
- Gell-Mann-Oakes-Renner relation, 347
- Generalized projective group
 - generators, 243
 - infinite dimensional, 243
 - Lie algebra, 243
 - oscillator representation, 243
- Generalized vertex operators, 239
- GGRT analysis, 262
 - anomaly, 268
 - canonical formalism, 264
 - classical algebras, 267
 - complete gauge fixing, 263
 - conserved angular momentum, 267
 - conserved momentum, 267
 - constraints, 265
 - covariant, 264
 - ends move with c , 263
 - Euler-Lagrange Eqns., 264
 - Hamiltonian, 266
 - independent variables, 266
 - light-cone parametrization, 265
 - mode expansion, 266
 - mode expansion of H , 267
 - Mode expansion of M^2 , 267
 - non-covariant, 264
 - non-covariant analysis, 268
 - normal modes, 267
 - normal ordering, 268
 - only transverse d.o.f, 263
 - Poisson brackets, 266
 - quantum Virasoro algebra, 268
 - solving the constraints, 266
 - $\tau = t$ gauge, 262
 - $\dot{X} \cdot X'$ condition, 263
- GGRT covariant analysis, 269
 - anomaly $\simeq D$, 272
 - covariant ETCR, 271
 - $1 \leq D \leq 26$, 272
 - ghosts, 271
 - mode expansion, 271
 - no ghosts proofs, 272
 - physical states, 272
 - quantization, 269
 - quantum constraints, 272
 - unit intercepts, 272
 - vs Goto analysis, 269
- GGRT non-covariant quantization, 269
 - anomaly $\simeq (D - 2)$, 269
 - $D=26$ issues, 269
 - DDF construction, 269
 - Lorentz covariance, 269
 - positive norm, 269
 - transverse d.o.f, 269
 - unit intercepts, 269
- Goldberger-Treiman relation, 164, 289, 347
 - chiral symmetry, 164
 - dispersion relation, 164
 - PCAC, 164
- Gold-foil expts., 5, 19
- Goldstone bosons, 285, 347
 - pions, 285
- Goldstone Theorem, 347
- Goto analysis, 269
 - classical aspects, 269
 - closure of algebra, 270
 - conflicts with anomaly, 271
 - conformal algebra, 270
 - Fourier vs Mode expansion, 270
 - no coordinate conditions, 269
 - No Euler-Lagrange eqns., 270
 - origin of constraints, 270
 - physical states, 271
 - strings, 269
- Grassmann variables, 353
- Group $SU(1,1)$, 242
 - generators, 242
 - Lie algebra, 242
 - non-compact, 242
- H**
- Haag's Theorem, 54
- Hadronic currents, 169
 - Cabibbo angle, 170
- Han-Nambu (HN) quarks, 26, 319, 325
 - breaking of colour, 323
 - breaking of flavour, 323
 - colour, 319
 - colour dependence of I_3 , Y , 320
 - flavour-colour mixing, 320
 - R-ratio, 325

tests for, 325

Harmonic oscillators

- fields from, 238
- N-point functions, 236
- propagators, 237
- vertex operators, 236
- zero modes, 238

Heat bath algorithm, 387

- Ising model, 387
- SU(2) LGT, 387

Heisenberg algebra, 236

Heisenberg representation, 66

Heisenberg S-matrix, 70

Heisenberg's S-matrix papers, 44

Higgs boson, 27

- discovery, 27

Hilbert transforms, 108, 127

Hole theory

- QFT equivalence, 71

Hypercharge Y, 318

Hyperfine structure, 10

Hyperspherical Legendre polynomials, 213

I

Interaction picture, 66

Interference model, 195

- double counting, 195

Internal conversion, 20

Invariants, 296

Ising models, 351

- duality, 351

Isospin transformations, 296

Isotopes, 6

Isotopic spin, 25, 323

- accidental symmetry, 323

J

Jacobi identities, 327

Jost-Lehmann-Dyson Theorem, 118

- integral representations, 120

K

KABRU, 396

Kallen-Lehmann representation, 88

Khuri-Kinoshita bounds, 167

Klein-Gordon Eqn., 13, 60, 84

- negative energy solns., 60

Koba-Nielsen variables, 225, 256

- anharmonic ratios, 226
- detailed constructions, 228
- duality constraints, 225
- projective invariance, 225

Kramers's mass renormaliation, 63

Kronig's Letter to Editor, 47

L

Lagrangian in QM, 69

Lamb Shift, 14, 61

- French-Weisskopf calculation, 64, 65
- Japanese approach, 65, 67
- Lamb-Kroll calculation, 64
- physics of, 63

Large order Feynman diagrams, 255

Lattice gauge theories

- abelian case, 365
- Alexander Polyakov, 353
- and path-integrals, 354
- confinement problem, 351
- $D = 1$ example, 360
- discretizations, 357
- elementary introduction, 359
- Euclidean formulation, 351
- Euclideanization, 356
- fermions, 354
- Haar measure, 371, 388
- Hamiltonian formulation, 351
- invariants, 370
- Jan Smit, 353
- Kenneth Wilson, 352
- link variables, 353, 368
- manifest gauge invariance, 353
- mapping to stat. mech, 353
- Monte Carlo simulations, 383
- no gauge fixing, 366
- non-abelian case, 366
- non-perturbative regularization, 351
- Osterwalder-Seiler work, 359
- path-integral, 371
- plaque action, 370
- polyakov loops, 370
- positive Hilbert space, 354
- renormalization, 388
- stat mech connection, 358
- statistical continuum limit, 361, 388
- transfer matrix, 359, 363
- Wegner-Wilson loops, 369
- Z_2 case, 351

Lehmann Ellipses, 108

- cos θ -analyticity, 122
- going beyond, 178
- large, 123, 152, 156
- properties, 117
- small, 123, 152, 156

Leptons, 24

Lie groups, 327

- adjoint rep., 327
- compact, 327

- lie algebras, 328
- structure constants, 327
- Links as holonomies, 353
- Link variables, 368
 - gauge transformations, 367
 - holonomy, 368
- Liouville action, 415
- Liquid drop model, 31
- London equation, 286
- Low energy theorems, 288
 - current algebras, 288
 - gravitational radiation, 288
 - Nambu–Lurie, 288
 - PCAC, 288
 - Weinberg, 288
- LSZ formalism, 78, 131, 134
 - non-perturbative S-matrix, 100
 - $\pi - N$ scattering, 95
 - proof of unitarity, 113
 - reductions, 87
 - R' -product, 110
 - and Symanzik, 142
 - T and R-products, 132
 - t-channel unitarity, 116
 - unitarity, 112
 - unretarded commutators, 102
- LSZ Theorem, 95
- Lüscher term, 392
 - LGT, 391
 - universality, 392
- LW EST, 403
 - boundary terms, 405
 - classical analysis, 409
 - higher order actions, 406
 - leading order, 403
 - open-closed duality, 406
 - R^{-5} order, 436
- M**
- Mandelstam double spectral rep., 158
- Mandelstam representation, 166, 169
 - complex angular momentum, 180
 - potential scattering, 179
 - vs dispersion relations, 182
- Mandelstam variables, 109, 144
- Manifest covariance, 65, 66
- Mass renormalization, 15
- Matrix mechanics, 8
- Maximal analyticity, 108
- Meissner effect, 25, 283, 286, 345
- Meson field theories
 - Bethe, 35
 - dispersion relations, 37
 - failures, 34
 - large coupling, 36
 - mixed meson theories, 35
 - S-matrix approach, 37
 - their failures, 31
- Meson–nucleon scattering, 163
 - dispersion relations, 163
 - S,P,D phase shifts, 164
- Metric conventions, 78
- Microcausality, 53, 85, 132, 155
 - essential?, 56
 - ETCR, 55
 - fermion fields, 55
 - field measurability, 55
 - field measurements, 74
 - for currents, 55
 - gauge fields, 55
 - Hamiltonian density, 55
 - RQFT, 53
- Minimal subtraction, 334
 - modified, 335
 - pros and cons, 335
- Monte Carlo simulations, 383
 - approach to equilibrium, 385
 - detailed balance, 386
 - Frobenius–Perron Theorem, 385
 - heat bath, 386
 - Markov processes, 385
 - Metropolis, 386
 - stochastic matrices, 385
- More quarks, 26
- Multilevel algorithm, 394
- Muon-decay, 23
 - parity violation, 22
- Muons, 23
 - vs pions, 23
- N**
- Nambu–Goldstone bosons, 25
- Nambu–Goto action, 258
 - boundary conditions, 261
 - classical aspects, 260
 - comparison with electrodynamics, 261
 - comparison with QED, 260
 - conserved momenta, 262
 - constraints, 261
 - coordinate conditions, 261
 - differences from GR, 259
 - Dirichlet and Neumann BC’s, 261
 - Euler–Lagrange Eqns., 261
 - Gauge-fixing, 260
 - geometrical, 258
 - induced metric, 258
 - invariances, 258
 - Poincaré invariance, 258, 262

- reparametrization invariance, 258
 - target space, 258
 - target space invariances, 262
 - Narrow resonances, 195
 - Negative energy solns., 69
 - backwards in time, 69
 - Negative frequencies, 104
 - Neutral pions, 36
 - exptal discovery, 36
 - spin, 36
 - theoretical discrepancies, 37
 - two photon decay, 36
 - Neutral weak boson, 26
 - Neutrino, 21
 - β -spectrum, 21
 - discovery, 22
 - masses, 22
 - Neutrons, 7, 11, 20
 - Neutron Stars, 7
 - Nielsen-Olesen strings, 346
 - dual strings, 346
 - Nielsen-Susskind action
 - conformal invariance, 258
 - gauge fixed nature, 258
 - Non-abelian gauge theories, 368
 - invariants, 368
 - Non-abelian theories
 - field content, 73
 - Non-forward scattering
 - dispersion relations, 144
 - Non-forward scattering, 144
 - Salam's approach, 144
 - Normal ordering, 244
 - N-point functions, 214
 - Bardakci-Ruegg, 222
 - Bootstrap condition, 222
 - degeneracy, 240
 - enlarged invariances, 229
 - factorizability, 240
 - Goebel-Sakita, 222
 - hooks, 222
 - in Koba-Nielsen variables, 226
 - issue of intercepts, 219
 - Jacobians, 230
 - multiperipheral configuration, 215
 - partitions, 222
 - 5 point function, 214
 - 6 point function, 220
 - Y.Nambu, 218
 - Nuclear democracy, 45
 - Nuclear forces, 23
 - attractive, 32
 - charge independence, 31
 - Lamb-Schiff theory, 35
 - observational, 31
 - range, 31
 - saturation, 31
 - short range, 32
 - spin-dependence, 31
 - strength, 31
 - theory, 32
 - Yukawa theory, 33
 - Nuclei, 11, 31
 - Odd-Even effect, 11
- O**
- Operator formalism, 233
 - dual resonance models, 233
 - more technical aspects, 241
 - N-point functions, 237
 - 4-point function, 237
 - 6-point function, 237
 - Operator product expansion, 342
 - Operators, 80
 - annihilation, 80
 - creation, 81
 - Optical Theorem, 189
 - Oscillator structure and dual models, 255
 - Overlapping divergences, 74
- P**
- Parastatistics, 319
 - Parity invariance, 135
 - Parity violation, 22
 - Partial wave amplitudes, 183
 - holomorphy, 183
 - Partial wave analysis, 176
 - asymptotic behaviours, 177
 - complex angles, 176
 - convergence, 176
 - large l behaviours, 178
 - Lehmann Ellipses, 176
 - Partial wave expansion, 175
 - Particles
 - identical, 77
 - indistinguishable, 13
 - Partitions of integers, 241
 - Parton distribution functions, 341
 - Parton model, 325, 341
 - Path integrals, 69, 354
 - configuration space, 354
 - and dual models, 445
 - extrinsic curvature, 447
 - large D analysis, 446
 - phase space, 354
 - saddle point approximations, 446
 - static potential, 446
 - subcritical strings, 445–448

Pauli Exclusion Principle, 8
 Pauli's Neutrino Letter, 21
 Pauli's New Qtm Number, 8
 Pauli-Villars Regularization, 68
 PCAC, 347
 Periodic table, The, 5
 Perturbative S-matrix, 72
 Phenomenological Lagrangeans, 304
 J.Schwinger, 304
 S.Weinberg, 304
 Photons, 73
 Physical regions, 104
 Physical sheet, 185
 Pion decay constant, 290
 Pion decay constant f_π , 170
 Pions, 23, 33
 parity, 36
 pseudoscalar, 33
 spin, 33
 vector theories, 35
 Pions as Goldstone bosons, 347
 Planar diagrams, 255
 Polonium, 19
 Polyakov approach, 275
 2-d quantum gravity, 275
 Hamiltonian formulation, 275
 string theory, 275
 subcritical strings, 275
 Polyakov loop correlators, 381
 spectral decomposition, 381
 Polyakov loops, 380
 Pomeranchuk Theorem, 167
 Pomeron, 191, 199
 Positrons, 13, 60, 69
 confused for protons, 13
 discovery, 13, 60
 Projective invariance, 223
 Haar measure, 226
 Koba-Nielsen variables, 223
 Möbius transformations, 226
 measure, 223
 Proton proton scattering, 32, 34
 Prout's atomic hypothesis, 5
 PS EST, 411
 Bosonic strings, 422
 for all D, 414
 higher orders, 416
 leading order, 412
 R^{-3} analysis, 418
 R^{-5} order, 437
 subcritical strings, 415
 Virasoro generators, 413, 418

Q

QCD
 advantages, 324
 background, 316
 colour confinement, 344
 current algebra, 324
 gauge fields, 328
 group theory, 327
 higher order beta Fn., 340
 infrared slavery, 324
 Lagrangian density, 329
 Λ parameter, 339
 local gauge invariance, 328
 log scaling violations, 342
 monopoles in, 345
 perturbative, 342
 static $Q\bar{Q}$ -potential, 343
 transformation rules, 329
 two-loop beta fn., 340
 Yang-Mills construction, 326
 QCD Λ , 339
 as integration onst., 339
 meaning of, 338
 QCD effective actions, 311
 QED, 59
 all orders, 62
 analyticity aspects, 75
 assessment of S-matrix, 74
 asymptotic series, 74
 causality and analyticity, 59
 constraints, 261
 divergences, 62
 Heisenberg S-matrix, 75
 lack of asymptotic states, 75
 Landau singularity, 336
 local gauge invariance, 326
 physical states, 261
 residual gauge invariance, 261
 S-matrix, 59, 73
 QFT, 43
 analyticity, 95
 asymptotic conditions, 79
 asymptotic states, 43, 79
 books on, 77
 causality principle, 53
 crossing symmetry, 103
 divergences, 43
 ETCR, 92
 from particles, 78
 fundamental length, 43
 Heisenberg-Pauli, 60, 66
 infrared problems, 79
 non-perturbative, 89
 non-renormalizable, 44

- normalizations, 86, 91
- 2-point function, 89
- relativistic (RQFT), 77
- retarded commutators, 100
- R-product, 101, 110
- self-energy, 61
- spectrum conditions, 90
- T-product, 99
- unitary transformations, 66
- vs Hole theory, 60
- Quadratic Casimirs, 341
- Quantized electromagnetic field, 62
 - and the atom, 62
- Quantum action principle, 69
- Quantum ambiguities, 45
- Quantum Chromodynamics (QCD), 26, 37
- Quantum Electrodynamics (QED), 12
- Quark confinement, 344
 - dual superconductivity, 344
 - lattice gauge theory, 344
 - monopoles, 344
- Quark dynamics, 322
 - Fritzsch-Gell-Mann, 323
 - gauging colour, 324
 - octet of gluons, 322
 - Pati-Salam, 323
- Quark models, 26, 318
 - acid tests, 322
 - Gell-Mann Zweig, 318
 - Han-Nambu, 322
 - π^0 -decay, 320
 - statistics problems, 319
- R**
- Radiating atoms, 14
- Radioactivity, 19
- Radium, 19
- Regge asymptotics, 186
- Regge Poles, 159, 179
 - bound states, 184
 - Coulomb scattering, 179
 - crossed channel, 187
 - direct channel, 187
 - high energy behaviour, 184
 - resonances, 185
- Regge trajectories, 194
- Relativistic hydrogen atom, 64
- Renormalizability, 15
- Renormalizability of SSB theories, 27
- Renormalization, 63
 - charge, 63
 - mass, 63
- Renormalization group, 330
 - Gell-Mann-Low analysis, 331
 - improving Pert. Th., 337
 - propagation function, 331
- QCD, 337
- QED, 335
- Renormalization of g_A , 169
 - Adler-Weisberger derivation, 169
 - and π^+p -cross sections, 172
 - dispersion relations, 169
- Rho mesons, 37
 - exptal discovery, 37
- RQFT
 - non-perturbative, 77
- Rubber string model, 255
- Running couplings, 330, 332
- S**
- Sakata model, 316
- Salam's generalized dispersion relations, 148
- Scaling violations, 342
- Scattering lengths, 306
 - pion-nucleon, 306
 - pion-pion, 306
- Schrödinger representation, 66
- Schwinger D=2 QED, 345
- Schwinger functions, 356
- Schwinger's canonical transformations, 65
- Schwinger's field perspective, 70
- Second Riemann Sheet, 186
- Self-energy, 61
- Shapiro-Virasoro (SV) model, 251
 - closed strings, 252
 - Koba-Nielsen variables, 252
 - massless spin-2, 253
 - oscillator representation, 251
 - tachyonic ground state, 251
 - vertex operators, 253
- Shelter Island, 62
- Shelter Island conference, 14
- Sliding scales, 330
- S-matrix, 15, 43, 189
 - analytical, 93
 - analyticity, 95
 - bounds, 167
 - causality, 128
 - classical mechanics, 43
 - crossing symmetry, 45, 103
 - J.A.Wheeler, 43
 - Lehmann integral representation, 121
 - non-perturbative, 100
 - QED, 15, 72
 - QFT, 44
 - strong interactions, 79
 - T-matrix, 97, 110

- unitarity, 107
 - S-matrix analyticity, 75
 - QED as test case, 74
 - Sommerfeld Puzzle, 12
 - Sommerfeld-Watson transform, 183
 - Beyond Lehmann Ellipses, 184
 - Spectral functions, 89
 - positivity, 90
 - Spectral representation, 89
 - Spin, 7
 - neutrons, 11
 - protons, 11
 - Spin-0 Lamb Shift, 63
 - Spin-orbit coupling, 9
 - Spin-statistics connection, 10, 56
 - Spontaneous Symmetry Breaking (SSB), 25, 285
 - Heisenberg, 285
 - Nambu, 285
 - States, 79
 - normalizations, 79
 - one particle stability, 81, 96, 100
 - vacuum stability, 81
 - Static $Q\bar{Q}$ -potential, 343, 379
 - and transfer matrices, 379
 - Arvis potential, 397
 - $D = 3$, 395
 - $D = 3$ SU(2), 395, 396
 - different schemes, 343
 - Hari Dass and Majumdar, 396
 - in $D \neq 26$, 398
 - infrared issues, 395
 - infrared singularities, 340
 - IR improvements, 397
 - Lüscher and Weisz, 394
 - Lüscher term, 396
 - perturbative, 395, 398
 - scaled second derivative, 396
 - two-loop QCD, 344
 - Wilson loops, 379
 - Statistical continuum limit, 361
 - analytical example, 363
 - correlation length scaling, 361, 389
 - phase transitions, 361
 - and weak coupling, 389
 - Stern–Gerlach expt., 7
 - Strange mesons, 37
 - Strangeness, 23, 316
 - associated production, 316
 - String actions, 256
 - String momentum, 422
 - higher orders, 422
 - String quantization, 269
 - String theory
 - Arvis quantization, 273
 - birth of, 255
 - classical, 256
 - closed strings, 256
 - extended objects, 256
 - fixed ends, 273
 - from dual models, 255
 - goto analysis, 269
 - longitudinal oscillations, 258
 - Nambu-Goto action, 258
 - Nielsen-Susskind action, 256, 257
 - open strings, 256
 - path integral quantization, 275
 - Polyakov action, 275
 - transverse excitations, 258
 - world-sheet coordinates, 256
 - world-sheet description, 255
 - Strong interactions, 288
 - effective description, 288
 - low energy theorems, 288
 - QCD the microscopic theory, 288
 - SU(1,1) generators, 242
 - oscillator representation, 241
 - Subtractions, 89
 - Sum rules, 197
 - Superconductivity, 283
 - BCS microscopic theory, 283
 - Ginzburg–Landau theory, 283
 - order parameter, 283
 - penetration depth, 287
 - real order parameter, 287
 - spontaneous symmetry breaking, 283
 - Superconvergence relations, 127, 130
 - general amplitudes, 191
 - π N -scattering, 190
 - Symmetries, 285
 - spontaneous breaking, 285
 - Systems with constraints, 270
- T**
- Tau-neutrino, 24
 - Tempered distributions, 159
 - Thomas Precession, 9
 - Toll’s logical equivalences, 128
 - Tomonaga’s field perspective, 70
 - Top quark, 27
 - Transition matrix, 43
 - quantum mechanics, 43
 - Translational invariance, 135
 - Transmutation of elements, 19
 - Transverse parity, 150
 - Tribute to Tomonaga, 65
 - Two neutrinos, 24
 - exptal confirmation, 24

U

Unified field theory, 44
 Unit intercept, 231
 symmetry enhancement, 231
 Universality of Fermi theory, 23
 Unphysical regions, 107
 Unstable particles, 81
 Upper Half Plane (UHP), 447

V

Vacuum polarization, 15, 61
 charge renormalization, 61
 divergences, 61
 Vacuum state, 80
 V-A theory, 22
 Vector pions, 33
 Veneziano formula
 Bootstrap conditions, 211
 crossing symmetry, 203
 daughter trajectories, 206
 even spacing of daughters, 206
 grand synthesis, the, 203
 harmonic oscillators, 233
 important properties, 205
 issue of daughters, 211
 manifest duality, 203
 mass relations, 206
 motivation, 204
 precise duality, 205, 206, 208
 Regge asymptotics, 205

 superconvergence relations, 206
 Virasoro algebra, 243
 anomaly and Jacobi identities, 243
 Visualizing QED, 69

W

Wave mechanics, 8
 Weak Hamiltonian, 22
 Weak neutral currents, 280
 Weak vector bosons, 24, 25
 Weyl connections, 429
 Weyl-weight compensators, 429
 Wick's Theorem, 73
 Wightman functions, 355
 Wilson loop-plaquette correlations, 382
 Wilson loops, 379
 area vs. perimeter laws, 380
 confinement criterion, 380
 Creutz ratios, 380
 spectral decomposition, 379
 W.Pauli's S-matrix critique, 45

Y

Yang-Mills fields, 25
 renormalizability, 25

Z

Zeeman Effect, 7
 anomalous, 7